The Dissertation Committee for Shenshen Yang
certifies that this is the approved version of the following dissertation:

Essays on Nonparametric and Semiparametric Identification and Estimation

Committee:

Jason Abrevaya, Co-Supervisor

Sukjin Han, Co-Supervisor

Brendan Kline

Haiqing Xu

Essays on Nonparametric and Semiparametric Identification and Estimation

by

Shenshen Yang

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2021

Dedicated to my parents,

my husband Nan,

and my soon-to-be-born daughter

Acknowledgments

My special thanks to my committee members, Jason Abrevaya, Sukjin Han, Brendan Kline and Haiqing Xu, for their invaluable support and guidance, both academically and emotionally during my whole graduate study. I am also grateful to Vasiliki Skreta, Caroline Thomas, Richard Murphy, Daniel Ackerberg and Dean Spears for their helpful suggestions on my dissertations. Thanks to Xun Li, Liyong Xu, Lu Chen, Jack Porter, Xiaoxia Shi, and David Knight, who gave me important advice and guidance on my academic path.

My gratitude also goes to my friends and colleagues Jiangang Zeng, Eric H Schulman, Qingsong Pan, Jessie E Coe, James Brand and Danyang Zhao for their support and valuable feedback at different stages of my research. I would also like to thank my supportive friends Xue Li, Shaofei Jiang, Chan Yu, Nir Eilam, Sangwoo Choi for their help during the past five years.

Finally, I want to thank my parents and grandfather, who encouraged me into graduate study and are always there for me. Thanks to my loving husband for his limitless support and encouragement.

Essays on Nonparametric and Semiparametric Identification and Estimation

by

Shenshen Yang, Ph.D.
The University of Texas at Austin, 2021

Co-Supervisors: Jason Abrevaya
Sukjin Han

This dissertation consists of three chapters in econometric theory, with a focus on identification and estimation of treatment effect in semi-parametric and nonparametric models, when there exists endogeneity problem. These methods are applied on policy and program evaluation in health and labor economics.

In the first chapter, I examine the common problem of multiple missing variables, which we refer to as multiple missingness, with non-monotone missing pattern and is usually caused by sub-sampling and a combination of different data sets. One example of this is missingness in both the endogenous treatment and outcome when two variables are collected via different stages of follow-up surveys. Two types of dependence assumptions for multiple missingness are proposed to identify the missing mechanism. The identified missing

mechanisms are used later in an Augmented Inverse Propensity Weighted moment function, based on which a two-step semiparametric GMM estimator of the coefficients in the primary model is proposed. This estimator is consistent and more efficient than the previously used estimation methods because it includes incomplete observations. We demonstrate that robustness and asymptotic variances differ under two sets of identification assumptions, and we determine sufficient conditions when the proposed estimator can achieve the semiparametric efficiency bound. This method is applied to the Oregon Health Insurance Experiment and shows the significant effects of enrolling in the Oregon Health Plan on improving health-related outcomes and reducing out-of-pocket costs for medical care. The method proposed here provides unbiased and more efficient estimates. There is evidence that simply dropping the incomplete data creates downward biases for some of the chosen outcome variables. Moreover, the estimator proposed in this paper reduced standard errors by 6-24% of the estimated effects of the Oregon Health Plan.

The second chapter is a joint work with Sukjin Han. In this chapter, we consider how to extrapolate the general local treatment effect in a nonparametric setting, with endogenous self-selection problem and lack of external validity. For counterfactual policy evaluation, it is important to ensure that treatment parameters are relevant to the policies in question. This is especially challenging under unobserved heterogeneity, as is well featured in the definition of the local average treatment effect (LATE). Being intrinsically local, the LATE is known to lack external validity in counterfactual environments. This

chapter investigates the possibility of extrapolating local treatment effects to different counterfactual settings when instrumental variables are only binary. We propose a novel framework to systematically calculate sharp nonparametric bounds on various policy-relevant treatment parameters that are defined as weighted averages of the marginal treatment effect (MTE). Our framework is flexible enough to incorporate a large menu of identifying assumptions beyond the shape restrictions on the MTE that have been considered in prior studies. We apply our method to understand the effects of medical insurance policies on the use of medical services.

In the third chapter, I investigate the partial identification bound for treatment effect in a dynamic setting. First, I develop the sharp partial identification bounds of dynamic treatment effect on conditional transition probabilities when the treatment is randomly assigned. Then I relax the randomization assumption and gives partial identification bounds, under a conditional mean independence assumption. Using MTR and MTS assumptions, this bound is further tightened. These bounds are used on estimating labor market return of college degree in a long term, with data from NLSY79.

Table of Contents

List of Tables

List of Figures

Chapter 1

IV Models with Missing Outcomes and Treatments, with Application to Oregon Health Insurance Experiment

1.1 Introduction

Missingness in multiple variables is common in practice. This phenomenon is caused by various reasons, including sub-sampling and a combination of different data sets; a typical example is missingness in both the endogenous treatment status and the outcome. This problem usually appears in empirical works but has not drawn enough attention. Researchers often observe missingness on both the endogenous treatment and outcome variables in a data set collected by surveys in both observational studies and field experiments. One approach that is regularly used by practitioners is to drop all observations in incomplete data; this is referred to as the complete case (CC) analysis. It is well known that the CC approach creates an inefficient estimator, and when the missing mechanism depends on endogenous variables, it is also biased (Roderick JA Little and Donald B Rubin (2002); Li Qi and Yanqing Sun (2014)). This paper proposes consistent and more efficient estimators with multiple missingness, which allows for the missing mechanism to be endogenous.

When there are two missing variables, the missing patterns can be divided into monotone missingness and strict non-monotone missingness; both are examples of a general non-monotone missing pattern.[1] Monotone missingness has been thoroughly studied in the literature (Anastasios Tsiatis (2007); Jean-Louis Barnwell and Saraswata Chaudhuri (2018); Saraswata Chaudhuri (2020)). It refers to the situation where the missingness of one indicates the missingness of the other variable. In our framework, this means that missingness in treatment implies missingness in the outcome. The monotone missing pattern can be caused by data attrition when once the survey participants drop out, they never return. In survey-collected data, missing variables can be caused by a broad range of reasons; as a result, strict non-monotone missingness often appears in a data set, which means that researchers can observe the outcome status for some observations, despite the missing treatment status. Survey data frequently suffers from strict non-monotone missingness; therefore, this paper focuses on strictly non-monotone missingness and discusses how to generate the proposed approach for more general missing patterns.

There are two sources of endogeneity in the model. First, the full model with no missing values is endogenous because of the endogenous treatment variable. Second, the missingness can be endogenous and correlated with latent variables in the model. The endogeneity in the full model is often addressed using an exogenous instrumental variable, which is correlated with the

[1]The general non-monotone missing pattern also includes univariate missingness on the treatment variable or the outcome variable.

endogenous regressor, but does not directly affect the outcome. This strategy has been widely used in the two-stage least squares (2SLS) estimation procedure, the generalized method of moments (GMM) model, and a more general nonparametric IV model (Henri Theil (1971); Joshua D Angrist and Guido W Imbens (1995*b*); Christopher F Baum, Mark E Schaffer and Steven Stillman (2003); Whitney K Newey and James L Powell (2003); Whitney K Newey (2013)). The other source of endogeneity is from selective missingness. Prior studies have shown that if the Missing at Random (MAR) assumption is satisfied—meaning that the missing mechanism is independent with missing values conditional on fully observed variables—the endogeneity of missingness can be solved (Philip E Cheng (1994); Shaun Seaman, John Galati, Dan Jackson and John Carlin (2013)). However, the non-monotone missing pattern complicates this assumption, because simultaneous dependence between multiple missingness creates challenges in identifying the joint missing mechanism.

When data are collected sequentially (i.e., the treatment variable is collected before the outcome variable), the simultaneous dependence relationship is avoided. This happens when the realization of outcome status takes time and is collected via a later follow-up survey. Motivated by the sequential data collection process, we propose two sets of identifying assumptions to identify the missing mechanism. The first is the MAR assumption, which assumes that missingness in the treatment and the outcome are independent with unobserved values conditioning on the fully observed variables, but this assumption does not allow the missing mechanism to depend on partially observed

variables. The alternative identifying assumption allows for the later-stage missingness (i.e., missingness on the outcome) to rely on partially observed variables in the previous stage (i.e., the partially observed treatment variable), which we describe as the sequentially updating feature of the missing mechanism; due to this feature, we name the assumption the Sequential MAR (SMAR) assumption.

Under either identifying assumption, the missing mechanism (i.e., the propensity of each missing pattern) is identified and utilized in an augmented inverse propensity weighted (AIPW) GMM estimator for the primary model coefficients. The moment function is composed of an inverse propensity weighted (IPW) moment function and an augmentation term chosen to make full use of the observed data and achieve higher efficiency; this function is equivalent to a two-step backward AIPW imputation approach, in which the missing outcome is initially imputed, followed by the imputation of the missing treatment.

The estimation strategy depends on a first-stage estimation of nuisance parameters, which are the propensities of missing patterns and models for incomplete data; we use sieve estimation for the first stage nuisance parameters. Even though AIPW has been shown to maintain double robustness and semiparametric efficiency in many cases (e.g., James M Robins, Andrea Rotnitzky and Lue Ping Zhao (1994); Daniel O Scharfstein, Andrea Rotnitzky and James M Robins (1999); Adam N Glynn and Kevin M Quinn (2010); Xiaohong Chen, Han Hong, Alessandro Tarozzi et al. (2008); Matias D Cattaneo (2010)), the desired properties usually fail under non-monotone missingness (e.g., Tsi-

atis (2007); Saraswata Chaudhuri and David K Guilkey (2016); Shaun R Seaman and Stijn Vansteelandt (2018); BaoLuo Sun and Eric J Tchetgen Tchetgen (2018)). We show that under the strong MAR assumption, the double robustness and semiparametric efficiency properties are maintained, and this result is consistent with the findings of Chaudhuri and Guilkey (2016). The estimator is robust under the SMAR assumption, as long as the missing mechanism is correctly specified; as a result, the asymptotic variance is affected by the first stage missing mechanism estimation. We provide asymptotic variance for the estimator under the SMAR assumption and show that the estimator become more efficient than previously used estimators by incorporating the incomplete data.

The AIPW-GMM approach is used in the Oregon Health Insurance Experiment to estimate the effect of enrolling in the Oregon Health Plan (OHP) on health-related outcomes. The endogenous treatment variable and outcomes are collected via different-stage follow-up surveys. There are two reasons behind the missingness: non-response to surveys and non-response to treatment-and-outcome-related questions among survey responders. There exist participants who did not respond to the first follow-up survey but responded to the final survey, and some participants did not answer questions on treatment status; therefore, we observe non-monotone missing patterns. The data shows evidence that missing mechanisms of outcome variables are correlated with the endogenous treatment variable. As a result, the CC analysis yield biased estimation results. The regression results show significant effects of the

OHP on reducing out-of-pocket costs for medical care, reducing the number of days when physical health is not good, and improving self-evaluated health conditions; these results suggest that CC estimators tend to overestimate the effect, with downward biases for out-of-pocket costs and self-evaluated health conditions outcomes, while the IPW estimator results are closer to those of the AIPW estimators. Furthermore, of the three different estimation strategies, the AIPW-GMM estimators achieve the smallest standard error for all estimated coefficients. For the estimated effects of OHP, the AIPW approach reduces the standard error by up to 24%. These results are consistent with the findings in the Monte Carlo simulation.

This paper contributes to the literature in the following aspects. First, it considers the problem of missing endogenous treatment variables and outcome variables, which lacks adequate attention. Related literature includes the partial identification approach developed by Joel L Horowitz and Charles F Manski (2000), when there exist multiple missing covariates and outcomes in randomized experiments. By making assumptions that can be plausible in many cases of program evaluation, this paper proposes a way to point identify the missing mechanism and to use it in the construction of a consistent and more efficient estimator. Second, this study contributes to the literature on non-monotone missingness (James M Robins and Richard D Gill (1997); Ahmed M Gad (2011); Sun and Tchetgen Tchetgen (2018); Eric J Tchetgen Tchetgen, Linbo Wang and BaoLuo Sun (2018); BaoLuo Sun, Neil J Perkins, Stephen R Cole, Ofer Harel, Emily M Mitchell, Enrique F Schisterman and

Eric J Tchetgen Tchetgen (2018); Chaudhuri and Guilkey (2016)) by adapting the sequentially updating feature of the missing mechanism[2] and allowing the missing mechanism to depend on a partially observed variable in a non-monotone missing pattern. Finally, this paper provides sufficient conditions under which the closed-form efficient influence function is available under non-monotone missingness.

The rest of this paper is organized as follows. Section 1.2 introduces the model; Section 1.3 missing mechanism and the assumptions and gives the identification result; Section 1.4 proposes the AIPW moment condition and GMM estimator based on the assumptions introduced in the previous sections; Section 1.5 discusses the asymptotic properties of the AIPW estimator; Section 1.6 extends the estimator to a more general set-up and shows the possibility of developing the current framework for more than two missing variables; Section 1.7 illustrates the performance of the AIPW-GMM estimator through the Monte Carlo simulation results; Section 1.8 offers an empirical example using the Oregon Health Insurance Experiment Data; and Section 1.9 will provide our conclusions.

1.2 Model and Assumptions

We consider the following model:

[2]Similar sequential feature has only been applied when the missing pattern is monotone (Chaudhuri (2020)).

$$Y_i = g(D_i, X_i; \beta) + \epsilon_i \tag{1.2.1}$$

where Y_i denotes the outcome of individual i and is missing for some observations; and D_i is a partially observed endogenous treatment variable. For purposes of this study, we only consider one partially missing endogenous treatment variable. X_i is a vector of K fully observed regressors, and we do not exclude the possibility that X_i contains endogenous variables.

We are interested in consistently estimating the parameter vector β, and we assume that there exists a vector of valid instrument variables Z_i for $[D_i, X_i]$ with $d_Z \geq d_D + d_X$, such that

$$E[Z_i \epsilon_i] = 0 \tag{1.2.2}$$

$$cov(Z_i, [D_i, X_i]) \neq 0 \tag{1.2.3}$$

Equations 1.2.2 and 1.2.3 are the exogeneity and relevance conditions, under which Z_i is a valid instrument variable. We assume that Z_i plays a role in determining $[D_i, X_i]$. We do not place structural restrictions on the first-stage model; therefore, the regressors can be either discrete or continuous.[3] In

[3]If more structures are added to the first stage, we could obtain additional moment conditions; as an example, a linear specification on D_i such that:

$$D_i = Z_i'\gamma + u_i$$

provides an extra moment condition besides Equation 1.2.4:

the full model, the moment condition follows directly from the exogeneity of Z_i:

$$E[Z_i(Y_i - g(D_i, X_i; \beta))] = 0 \tag{1.2.4}$$

Under the standard regularity conditions[4] for g, the parameter value of interest β^0 is identified by the moment conditions:

$$\beta = \beta^0 \text{ if and only if } E[Z_i(Y_i - g(D_i, X_i; \beta))] = 0 \text{ for } \beta \in \mathcal{B}$$

We consider an environment where the instrument variables Z_i and fully observed covariates X_i are fully observed, but the endogenous treatment D_i and the outcome variable Y_i are partially missing. We use R_i^D and R_i^Y to indicate observing D_i and Y_i, such that

$$R_i^D = \begin{cases} 1 & D_i \text{ is observed} \\ 0 & D_i \text{ is unobserved} \end{cases} \tag{1.2.5}$$

$$R_i^Y = \begin{cases} 1 & Y_i \text{ is observed} \\ 0 & Y_i \text{ is unobserved} \end{cases} \tag{1.2.6}$$

$$E[Z_i u_i] = 0$$

This extra information potentially helps to improve the precision and efficiency of the estimator (Rai 2020).

[4]The regularity conditions can be found in KW Newey and Daniel McFadden (1994).

In the following section, we will discuss the missing patterns and dependence between R^D, R^Y, and the other variables.

1.3 Patterns of Missingness

We first capture the missing mechanism and give two sets of assumptions to identify the propensity of missingness. The commonly used assumptions for univariate missing values are the Missing Completely at Random (MCAR) and the Missing at Random (MAR) assumptions. The MCAR assumes that missingness is independent from any variable in the data set, and MAR assumes that missingness is independent from the missing values conditioning on a set of observables.[5] Missing mechanisms that fail the MCAR and MAR assumptions are called Missing Not at Random (MNAR); these allow the missingness to be correlated with the unobserved missing values, and therefore creates difficulties in identifying the missing mechanism. Extra parametric assumptions are usually imposed when the missing pattern is MNAR (Scharfstein, Rotnitzky and Robins (1999); Rotnitzky Andrea, Daniel Scharfstein, Ting-Li Su and James Robins (2001)).

The existence of multiple partially missing variables complicates the missing mechanism. Based on the features of the missing patterns, prior studies have described multiple missing patterns as monotone missingness versus non-monotone missingness. Monotone missingness is defined as the situation

[5]These two assumptions can be seen as strong ignorability and conditional ignorability assumptions.

where missingness happens gradually, and missingness in one variable indicates missingness in other variables. In our framework, it is implied that Y_i is missing when D_i is missing and vice versa. We focus on the former possibility, since it is a more reasonable scheme compared to the latter. The missing mechanism is formally defined as monotone if $(1 - R_i^D)R_i^Y = 0$ almost surely.

For monotone missingness, each stage can be seen as a subsample of the last step, and the missingness depends on the fully observed variables from the previous stage. One reason behind this pattern is data attrition; when the program participants choose to drop out of the survey, they do not return to the following surveys of their outcomes. Another example is data composed of multi-phase sampling, from which the researchers choose a subsample of the previous phase to collect information on, due to budget constraints and survey design.

Non-monotone missingness is a more general pattern that allows for the possibility of observing Y_i while D_i is missing; it is also referred to as general missingness (Stef Van Buuren (2018)), and includes well-discussed univariate missingness,[6] monotone missingness, and strict non-monotone missingness. Non-monotone missingness arises in various scenarios other than data attrition, including general survey non-responses, panel studies in which par-

[6]Discussions on univariate missingness on the regressor can be found in Roderick JA Little (1992), Jason Abrevaya and Stephen G Donald (2017), Christoph Breunig and Peter Haan (2018); Lu Wang, Andrea Rotnitzky and Xihong Lin (2010), Rolf HH Groenwold, A Rogier T Donders, Kit CB Roes, Frank E Harrell Jr and Karel GM Moons (2012), Jason Abrevaya (2019) studied on methods dealing with missing outcome variables.

ticipants drop out but return in later surveys, and general sub-sampling in each stage. In Figure 1.1, we illustrate the missing patterns included in non-monotone missingness; among the patterns, strict non-monotone missingness is the most frequently seen and interesting case, so we will focus on this case.

With a non-monotone missing pattern, the previous MAR assumption is difficult to justify (James M Robins (1997); Robins and Gill (1997); Little and Rubin (2002); Sun and Tchetgen Tchetgen (2018))[7], and it creates challenges in identification. The challenge was first introduced by Robins and Gill (1997) when they considered a case where two partially missing variables exist and showed that MAR assumption implicitly implies the MCAR assumption in a logistic model; when there are other fully observed variables, the traditional MAR becomes a stronger assumption of conditional ignorability for the same observables. In our framework, because we have two potentially missing variables in the environment, we can observe four missing patterns in the data. We use M to denote the four patterns and introduce difficulties in point identifying the distribution of the missing patterns. We suppress the index i for the following arguments:

[7]The statistical literature has shown that for commonly used estimation process, it is difficult to include all features allowed by the previous MAR assumption. In this paper, we focus more on the challenge in identification, and show some features allowed by the previous MAR assumption must be excluded to identify the missing mechanism.

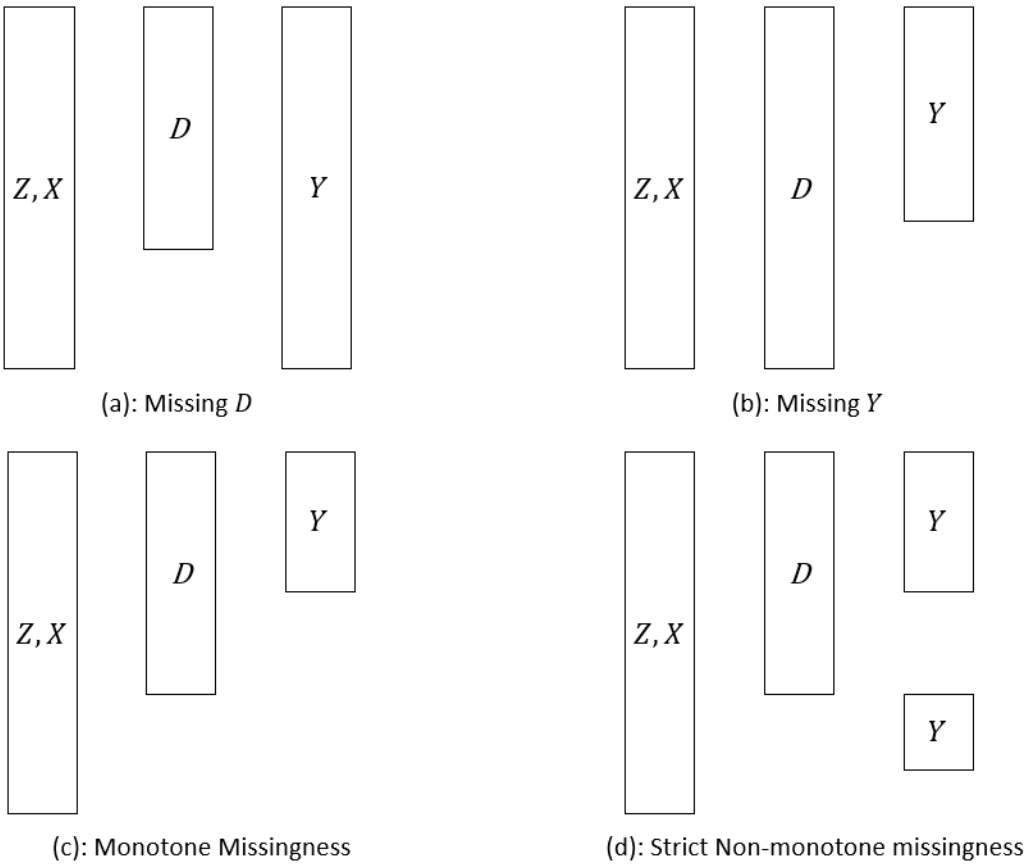


Figure 1.1. Examples of monotone missing patterns

$$M = M_1 : \text{Observe both D and Y}$$

$$M = M_2 : \text{Observe D but not Y}$$

$$M = M_3 : \text{Observe Y but not D}$$

$$M = M_4 : \text{Observe neither D nor Y}$$

The MAR assumption assumes mean independence between missingness and missing values conditioning on the observables under each missing pattern and can be written as:

$$\begin{aligned}
\Pr[M = M_1|Z, X, D, Y] &= \Pr[M = M_1|Z, X, D, Y] \\
\Pr[M = M_2|Z, X, D, Y] &= \Pr[M = M_2|Z, X, D] \\
\Pr[M = M_3|Z, X, D, Y] &= \Pr[M = M_3|Z, X, Y] \\
\Pr[M = M_4|Z, X, D, Y] &= \Pr[M = M_4|Z, X]
\end{aligned} \tag{1.3.1}$$

The first equation is intentionally redundant to show that no information is lost under the pattern M_1. Equation 1.3.1 implies Equation 1.3.2[8]:

$$\begin{aligned}
E[R^D|Z, X, D, Y, R^Y] &= E[R^D|Z, X, R^Y Y, R^Y] \\
E[R^Y|Z, X, D, Y, R^D] &= E[R^Y|Z, X, R^D D, R^D]
\end{aligned} \tag{1.3.2}$$

[8]Derivation is omiited, and the intuition behind is that each missing pattern depends on the observed variables under that pattern. Therefore, when $R^Y = 1$, $\Pr[R^D = 1|Z, X, D, Y, R^Y = 1]$, as a part of the missing mechanism, depends on Y; and when $R^Y = 0$, $\Pr[R^D = 1|Z, X, D, Y, R^Y = 1]$ does not depend on Y anymore. Therefore, $E[R^D|Z, X, D, Y, R^Y] = E[R^D|Z, X, R^Y Y, R^Y]$; same argument can be applied on $E[R^Y|Z, X, D, Y, R^D]$.

One way to model Equation 1.3.1 is using the threshold crossing model. To illustrate how the simultaneous relationship affects identification, we can rewrite the assumption in Equation 1.3.2 into the following simultaneous binary model[9]:

$$\begin{aligned} R^D &= 1\left[f(Z, X, R^Y Y, R^Y) \geq \mu\right] \\ R^Y &= 1\left[g(Z, X, R^D D, R^D) \geq \nu\right] \end{aligned} \tag{1.3.3}$$

for some unknown functions f and g, with μ and ν as uniformly distributed latent variables. The simultaneous feature in the binary model creates a problem in the identification of $\Pr\left[R^D = 1, R^Y = 1 | Z, X, D, Y\right]$. A similar model has been widely discussed in the entry game model, and there is a well-known challenge to identifying joint distribution of the simultaneous bivariate model. Prior studies provided partial identification strategies for the simultaneous binary model (e.g., Elie Tamer (2003*a*); Federico Ciliberto and Elie Tamer (2009*a*); Jorge F Balat and Sukjin Han (2018)) and point identification under additional restrictions. Even though identifying the missing mechanism is an important step to identifying the main model, that is not our primary interest. As an intermediate step, we want to avoid either partial identification or too-complex structural assumptions. Therefore, we provide two sets of novel ignorability-type assumptions that could be satisfied under many settings and rule out the simultaneity, and we provide justifications for these assumptions.

[9]The equivalence between Equations 1.3.2 and 1.3.3 follows from the property of binary variables.

These assumptions are used to point identify the missing mechanism.

1.3.1 Missing at Random Assumption

The first assumption is a stronger version of the MAR used in the univariate missingness literature by assuming the ignorability of multiple missingness and conditioning on a common set of fully observed variables.[10]

Assumption MAR.

$$R^D \perp\!\!\!\perp (D, Y)|Z, X \qquad (1)$$

$$R^Y \perp\!\!\!\perp (D, Y)|Z, X, R^D \quad (2)$$

We refer to this assumption as the MAR assumption for convenience, following the name used in prior studies. Chaudhuri and Guilkey (2016) applied analogous assumptions on non-monotonically missing instrumental variables and assumed the missingness was independent with unobserved values conditioning on fully observed variables.

If there are no endogenous variables in X, the MAR assumption can be interpreted as exogenous missingness; this assumes the joint missingness is independent with unobserved values conditioning on fully observed exogenous covariates. Missingness is determined by instrument variables (e.g., random assignment) and personal characteristics (e.g., age, education, distance to research center).

[10]Prior studies have implicitly shown that the MAR assumption in a univariate missingness setting is justified under the strong version of MAR; this can be displayed using the example from Sun and Tchetgen Tchetgen (2018), except for minor differences.

Example 1. *In field experiments, experimenters collect data via face-to-face visits. Missingness is believed to be caused by missed experimenter visits, either accidentally or by design, instead of self-selection of the participants. Therefore, missingness on D and Y depends on survey participant characteristics (e.g., location, gender, age, etc.) and instrument variables (e.g., random assignment into treatment group), instead of the treatment and outcome status.*

1.3.2 Sequential Missing at Random Assumption

The MAR assumption introduced above does not allow the missingness to depend on partially observed variables, and it is unclear how R^Y depends on R^D. We therefore propose a different assumption that allows for dependence between partially observed D and the missingness mechanism on Y, which occurs after the realization of R^D.

When missingness happens sequentially, the missing process has the "dynamic updating" feature introduced in Chaudhuri (2020), and this feature is applied to the monotone missing pattern. Despite the non-monotone pattern, this dynamic feature is still allowed in our framework. We can therefore make a Sequential MAR (SMAR) assumption that missingness on D is independent with unobserved values conditional on the fully observed variables, and missingness on Y is independent with the unobserved values conditional on fully observed variables and the partially observed D.

The SMAR assumption is formalized below:

Assumption SMAR. [11]

$$R^D \perp\!\!\!\perp (D, Y)|Z, X \tag{1}$$

$$R^Y \perp\!\!\!\perp (D, Y)|Z, X, R^D D, R^D \tag{2}$$

and (2) *is equivalent to*

$$R^Y \perp\!\!\!\perp ((1 - R^D)D, Y)|Z, X, R^D D, R^D \tag{2'}$$

The SMAR assumption is weaker than the MAR in the sense that it allows R^Y to be correlated with D; on the other hand, R^Y can be correlated with Y through the correlation with D. Therefore, it assumes that R^Y is independent with Y conditioning on the fully observed variables assumed in the MAR assumption, as well as $R^D D$. One important feature allowed by the SMAR assumption is the dependence between R^Y and $R^D D$, unlike the MAR assumption; and the other crucial feature assumed in SMAR is that D does not affect R^Y if D is not observed. In the following, we will provide an example of when these features hold.

Example 2. *In survey-collected data sets, missingness can be due to both exogenous attrition and endogenous self-selection. For the self-selected missingness, we make an extra assumption that survey participants are honest reporters to exclude distortion from misreporting. We assume the participants*

[11] Both the assumption MAR and the assumption SAMR can be relaxed to conditional mean independence.

choose to report if they have the information. Therefore, missing indicators R^D *and* R^Y *can be interpreted as indicators of awareness, or an effort to acquire the status of* D *and* Y*.*

Knowledge of treatment status D_i *can be caused by participants' characteristics, as well as the instrument variables. Age, living area, and education level can all affect participants' motivation to acquire their treatment status. The instrument variables can also affect* R^D*. One classic instrument variable used in the field-experiment literature is the random assignment of treatment, with non-perfect compliance as the instrumented endogenous variable. Researchers are likely to visit participants in the treatment group more than the control group, which might cause lack of awareness of treatment status for the control group. Moving to the next stage, if participants have no knowledge of their treatment status (i.e.,* $R^D = 0$*),* D *does not enter their information set and thus does not affect their motivation to attain knowledge about their outcome status (i.e.,* D *does not affect* R^Y*). This explains the assumed non-correlation between* R^Y *and* D *when* $R^D = 0$*.*

On the other hand, for participants who already have information about D*, this enters their information set and plays a role in determining whether or not they learn their outcome status. This is consistent with the feature that* R^Y *depends on* D *when* $R^D = 1$*, which is allowed in the SMAR assumption but not allowed in MAR assumption; as a result, the missing mechanism diverges for those with and without information about* D*. One example of this is that people with health insurance tend to care more about their health status because*

it affects their premium in the next billing cycle.

Figure 1.2 illustrates the sequential missing procedure. We use I_1 to denote the initial information set, which includes the fully observed variables. Variables in I_1 play roles in determining R^D; when $R^D = 1$, D is contained in the information set I_2, and when $R^D = 0$, D is not included in I_2. I_2 contains the variables determining R^Y. The missing mechanisms of Y diverge at I_2, and D only affects R^Y when $R^D = 1$. Therefore, the missing mechanism at each stage diverge across different observed variables in the previous stage.

Assumptions MAR and SMAR differ when it comes to the hypothetical missing mechanism on R^Y. We define p_d and p_y as:

$$p_d \equiv \Pr[R^D|Z, X, D, Y] = \Pr[R^D|Z, X] \tag{1.3.4}$$

$$p_y \equiv \Pr[R^Y|Z, X, D, Y, R^D] = \Pr[R^Y|\mathcal{C}, R^D] \tag{1.3.5}$$

The second equality in Equation 1.3.4 holds under either assumption MAR or assumption SMAR. $\mathcal{C}$ represents the conditioning variables for ignorability of R^Y; under the assumption MAR, $\mathcal{C} = (Z, X)$; and under the assumption SMAR, $\mathcal{C} = (R^D D, Z, X)$[12]. The distribution of missing patterns derives from the multiplication of p_d and p_y. For each missing pattern M_m, $m = 1, 2, 3, 4$, the probability of observing it is equal to

[12]For the assumption MAR, $\mathcal{C} = (R^D D, Z, X)$ can also be used without affecting consistency and asymptotic variance, similar to a statement that was made in Qi and Sun (2014).

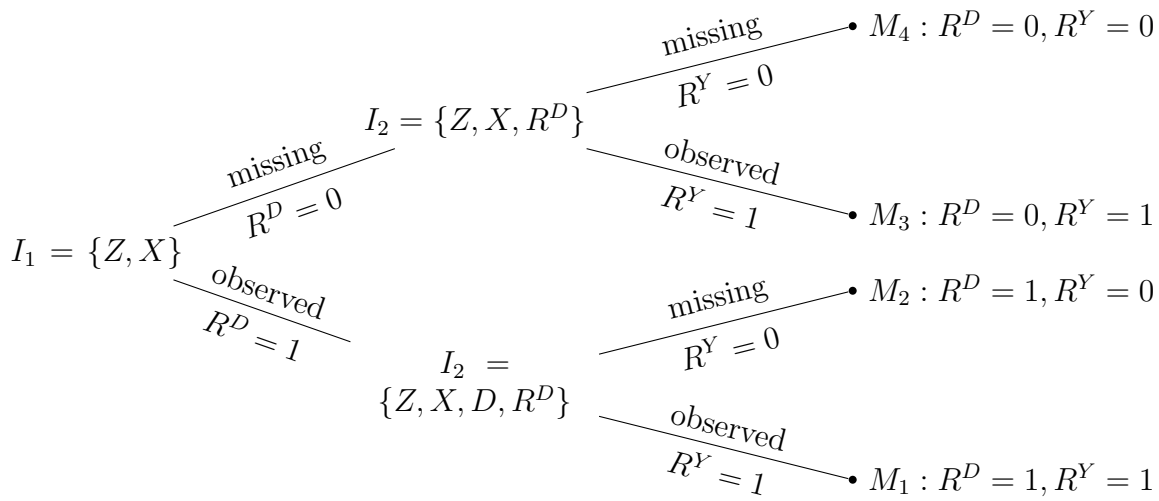


Figure 1.2. Sequential Missing Mechanism

$$\Pr[M_m|Z, X, D, Y] = \Pr[R^Y = M_m^Y|\mathcal{C}, R^D] \times \Pr[R^D = M_m^D|Z, X] \quad (1.3.6)$$

for M_m^D, M_m^Y being the corresponding value of R^D and R^Y under missing pattern M_m.

Proposition 1.3.1. *If either the Assumption MAR or the Assumption SMAR hold, the probability of missing patterns in 1.3.6 is identified.*

The identification result follows directly from selection on observables.

1.4 AIPW-GMM Estimator

The IPW estimator has been widely applied in the missing data literature (Paul R Rosenbaum and Donald B Rubin (1983); Jeffrey M Wooldridge

(2007); Shaun R Seaman and Ian R White (2013)). It reweights the sample and inflates the underrepresented subsample due to missingness to provide a consistent estimator. Compared to the imputation method (e.g., expectation-maximization, multiple imputations), the IPW estimator avoids making parametric assumptions on the model for an incomplete subsample and is easier to compute. The efficiency of the IPW estimator can be improved by adding an augmentation term, and this approach is referred to as the Augmented IPW (AIPW) approach. The AIPW estimator takes full advantage of the data set by incorporating dropped information in the IPW estimation into the augmentation term and usually achieves semiparametric efficiency bounds (Robins, Rotnitzky and Zhao (1994); Robins (1997); James R Carpenter, Michael G Kenward and Stijn Vansteelandt (2006); Tsiatis (2007); Chen et al. (2008); Glynn and Quinn (2010)). Moreover, the AIPW estimator has been shown to be robust in many cases where either the missing mechanism or the incomplete model is correctly specified. This property is called double robustness, and it creates the advantage that the first-stage nonparametric estimation of the missing propensity and incomplete model does not affect the efficiency of the AIPW estimator.

1.4.1 AIPW Moment Condition

First, we make an overlap assumption:

Assumption Overlap. *(a)* $p_d \in [\kappa_d, 1)$ *almost surely in* (Z, X)

(b) $p_y \in [\kappa_y, 1)$ *almostly surely in* $(\mathcal{C}, R^D)$ *for* $\kappa_d > 0$,$\kappa_y > 0$.

This assumption requires the possibility that each of the four missing patterns is positive (i.e., the strict non-monotone missing pattern).[13]

We further introduce the notation of probability from observing both the treatment and outcome, which is defined as:

$$p_{11}(X, Z, D) = \Pr[R^Y = 1|\mathcal{C}, R^D = 1] \times \Pr[R^D = 1|Z, X]$$

We construct a moment function $m_{aipw}(\beta)$:

$$m_{aipw}(\beta) = \frac{R^D R^Y}{p_{11}} Z(Y - g(D, X; \beta)) + \phi(Z, X, R^D, R^Y, R^D D, R^Y Y, \beta) \tag{1.4.1}$$

The moment function above is composed of an IPW moment function and an augmenting term ϕ, which is determined by fully observed variables and coefficient parameter β. We set the augmentation term to be:

$$\begin{aligned}\phi = &\left(\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_{11}}\right) Z\left[(Y - E\left[g(D, X; \beta)|Z, X\right]) - (E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right])\right] \\ &+ \left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right) Z\left[(E\left[Y|D, Z, X\right] - g(D, X; \beta)) - (E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right])\right] \\ &+ \left(1 - \frac{R^D R^Y}{p_{11}}\right) Z\left(E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right]\right)\end{aligned} \tag{1.4.2}$$

[13] A weaker version of the overlap assumption is introduced in Section 1.6, when a more general missing pattern is allowed.

The augmentation term is composed of observed variables, as well as two sets of nuisance parameters, the missing mechanism (p_d, p_y, p_{11}), and the imputed value for the unobserved model: $(E[g(D,X;\beta)|Z,X], E[Y|D,Z,X], E[Y|Z,X])$. Identification of the missing mechanisms was shown in the previous section. The assumptions of MAR or SMAR identifies the second set of nuisance parameters; under either assumption, these can be identified by[14]:

$$
\begin{aligned}
E[g(D,X;\beta)|Z,X] &= E\left[g(D,X;\beta)|Z,X,R^D=1\right] \\
E[Y|D,Z,X] &= E\left[Y|D,Z,X,R^D=1,R^Y=1\right] \\
E[Y|Z,X] &= E\left[Y|Z,X,R^D=0,R^Y=1\right]
\end{aligned}
$$

The moment function above can be interpreted as a two-step AIPW imputation. With full data, the moment function is:

$$m_{full}(\beta) = Z\epsilon = Z(Y - g(D,X;\beta))$$

The goal is to construct an unbiased estimator of ϵ under missingness on D and Y. Without partially missing D, the AIPW estimator for ϵ is written as:

$$\hat{\epsilon} = \frac{R^Y}{p_y}(Y - g(D,X;\beta)) + \left(1 - \frac{R^Y}{p_y}\right)(E[Y|D,Z,X] - g(D,X;\beta))$$

[14]Under Assumption MAR, $E[Y|Z,X]$ can also be identified by $E\left[Y|Z,X,R^D=1,R^Y=1\right]$ and $E\left[Y|Z,X,R^Y=1\right]$.

When D is partially missing, $g(D, X; \theta)$ and $E[Y|D, Z, X]$ are not fully observed as functions of D. We again use the AIPW strategy on the functions of D and obtain:

$$\begin{aligned}\hat{\epsilon} = \frac{R^Y}{p_y}\left[Y - \left(\frac{R^D}{p_d}g(D, X; \beta) + (1 - \frac{R^D}{p_d})E\left[g(D, X; \beta)|Z, X\right]\right)\right] \\ + \left(1 - \frac{R^Y}{p_y}\right)\left[\frac{R^D}{p_d}\left(E\left[Y|D, Z, X\right] - g(D, X; \beta)\right)\right. \qquad (1.4.3) \\ \left. + \left(1 - \frac{R^D}{p_d}\right)E\left\{E\left[Y|D, Z, X\right] - g(D, X; \beta)|Z, X\right\}\right] \qquad (1.4.4)\end{aligned}$$

Proposition 1.4.1. *If $\hat{\epsilon}$ is defined as in 1.4.4, $m_{aipw}(\beta) = Z\hat{\epsilon}$.*

The moment function is used to construct a moment condition, following the theorem below:

Theorem 1.4.1. *Suppose Assumption MAR or SMAR, and Assumption Overlap hold, $E\left[m_{aipw}(\beta)\right] = 0$ when $\beta = \beta^0$.*

The equality in this theorem will be used as the AIPW moment condition.

1.4.2 Estimation Strategy

The estimation procedure is composed of two steps. In the first step, we construct appropriate estimators for (p_d, p_y, p_{11}) and $(E[Y|D, Z, X], E[Y|Z, X]$, $E[g(D, X; \beta)|Z, X])$, and denote the estimators as $(\hat{p}_d,, \hat{p}_y, \hat{p}_{11})$ and $(\hat{E}[Y|D, Z, X]$,

$\hat{E}[Y|Z,X], \hat{E}[g(D,X;\beta)|Z,X])$. Estimation strategies on the nuisance parameters depend on the researcher's prior belief about model structures. If the correct specification of the nuisance parameter is not clear, nonparametric estimation could be applied to avoid models that are too restrictive. We will apply the series estimation strategy on the models, and we will show the corresponding asymptotic properties thereof in the following section.

After constructing the nuisance parameters, we plug them back into the GMM estimation equation and solve:

$$\hat{\beta}_{AIPW-GMM} = argmin_{\beta} \hat{m}_{aipw,i}(\beta)' \hat{W} \hat{m}_{aipw,i}$$

where

$$\hat{m}_{aipw} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i^D R_i^Y}{\hat{p}_{11,i}} Z_i \left(Y_i - g\left(D_i, X_i; \beta\right)\right) + \hat{\phi}_i$$

$$\begin{aligned}
\hat{\phi}_i = & \left(\frac{R_i^Y}{\hat{p}_{y,i}} - \frac{R_i^D R_i^Y}{\hat{p}_{11,i}} \right) Z_i \Big[\Big(Y_i - \hat{E}\left[g(D_i, X_i; \beta)|Z_i, X_i\right] \Big) - \\
& \Big(\hat{E}\left[Y_i|Z_i, X_i\right] - \hat{E}\left[g(D_i, X_i; \beta)|Z_i, X_i\right] \Big) \Big] \\
& + \left(\frac{R_i^D}{\hat{p}_{d,i}} - \frac{R_i^D R_i^Y}{\hat{p}_{11,i}} \right) Z_i \Big[\Big(\hat{E}\left[Y_i|D_i, Z_i, X_i\right] - g(D_i, X_i; \beta) \Big) \\
& - \Big(\hat{E}\left[Y_i|Z_i, X_i\right] - \hat{E}\left[g(D_i, X_i; \beta)|Z_i, X_i\right] \Big) \Big] \\
& + \left(1 - \frac{R_i^D R_i^Y}{\hat{p}_{11,i}} \right) Z_i \Big(\hat{E}\left[Y_i|Z_i, X_i\right] - \hat{E}\left[g(D_i, X_i; \beta)|Z_i, X_i\right] \Big)
\end{aligned}$$

is the estimator of the augmentation term ϕ and $\hat{W}$ is the estimator of chosen weight matrix W.

1.5 Asymptotic Properties

Though the AIPW estimator is known to have the efficient property in many cases, the efficiency usually fails with non-monotone missingness. One critical condition to produce closed-form semiparametric efficient bound is the independence of different missing indicators condition on the same set of variables (Chaudhuri and Guilkey (2016)); this condition implies that the asymptotic variances should be different under Assumption MAR and Assumption SMAR. Between these two conditions, Assumption MAR is more powerful, such that the AIPW-GMM estimator maintains the double robustness property and achieves the efficient semiparametric bound. Under the SMAR assumption, efficiency and double robustness fail, but it still has higher efficiency than the previously used estimators.

1.5.1 Asmptotic Properties

We make the standard assumptions for asymptotic normality in the GMM model:

Assumption M. *(1)* $(Z_i, X_i, D_i, Y_i, R_i^D D_i, R_i^Y Y_i, R_i^D, R_i^Y)$ *are i.i.d;*

(2)$E\left[m_{aipw}(\beta)\right]$ *is differentiable with respect to* $\beta \in int(\mathcal{B})$*;*

(3) Define $G(\beta) = \frac{\partial}{\partial \beta} E\left[m_{aipw}\right]$*,* $G(\beta)$ *has full rank at* $\beta = \beta^0$*;*

(4) V_{MAR} *and* V_{SMAR} *are bounded and positive semindefinite;*

We then derive the variance of the moment functions under different assumptions, and use these derivations to construct the asymptotic distribution

of the AIPW-GMM estimator. For a semi-parametric GMM estimation, the variance closely relates the to robustness of the estimator, thereby indicating how the first-stage nuisance parameters estimators affect the consistency of the primary model.

Under the Assumption MAR, the ignorabilities of R^D and R^Y depend on the same set of conditioning variables. The assumption is powerful enough that the double robustness property of AIPW estimator is maintained.

Theorem 1.5.1. *If Assumption MAR and Overlap hold,* $E\left[m_{aipw}(\beta)\right] = 0$ *when either* $(\hat{p}_d, \hat{p}_y, \hat{p}_{11})$*, or* $(\hat{E}\left[g(D, X; \beta)|X, Z\right], \hat{E}[g(D, X; \beta)|X, Z, Y], \hat{E}\left[Y|X, Z, D\right]$*,* $\hat{E}\left[Y|X, Z\right])$ *are correctly specified.*

As a direct result of double robustness, the first-stage estimators of the nuisance parameters do not affect the variance of the primary model.

Theorem 1.5.2. *Let* V_{MAR} *denote* $Var\left(m_{aipw}\right)$ *under assumption MAR and Overlap, then*

$$V_{MAR} = E\left[\frac{1}{p_{11}} Var\left(m_{full}|Z, X\right) + E\left[m_{full}|Z, X\right] E\left[m_{full}|Z, X\right]'\right] - \Delta$$

where

$$
\begin{aligned}
\Delta = Var & \left(\left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}} \right) \left(E\left[m_{full} | D, Z, X \right] - E\left[m_{full} | Z, X \right] \right) \right) \\
& + Var \left(\left(\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_{11}} \right) \left(E\left[m_{full} | Y, Z, X \right] - E\left[m_{full} | Z, X \right] \right) \right) \\
& + 2E \left[\left(1 - \frac{1}{p_y} - \frac{1}{p_d} + \frac{1}{p_{11}} \right) Cov\left(ZY, E\left[m_{full} | D, Z, X \right] | Z, X \right) \right. \\
& \left. + \left(\frac{1}{p_y} - \frac{1}{p_{11}} \right) Cov\left(Zg(D, X; \beta), ZY | Z, X \right) \right]
\end{aligned}
$$

The first term in V_{MAR} is similar to the Ω_β that was introduced in Chen et al. (2008), and Δ captures improvements in efficiency when we include the partially observed variables in the moment function.

Under the Assumption SMAR, the double robustness property no longer hold with non-monotone missingness. However, the estimator remains robust as long as the missing mechanism is correctly specified.

Theorem 1.5.3. *If Assumption SMAR and Overlap hold,* $E\left[m_{aipw}(\beta) \right] = 0$ *when* $(\hat{p}_d, \hat{p}_y, \hat{p}_{11})$ *is correctly specified.*

Proof of Theorem 1.5.3 follows directly from the proof of Theorem 1.4.1.

The double robustness fails under the SMAR assumption because the missing parts of the missing mechanism depend on partially missing variables. As a result, the consistency of some parameters in the first-step estimation of nuisance parameters affects the consistency and efficiency of the primary model and has a direct effect on the variance. Luckily, only the estimator $\hat{p}_{11}$ affects the consistency through the first component in the augmenting term,

and we can therefore construct a correction term following Whitney K Newey (1994).

Theorem 1.5.4. *Let V_{SMAR} denote $Var\left(m_{aipw}\right)$ under assumption SMAR and Overlap, then*

$$V_{SMAR} = E\left[\frac{1}{p_{11}} Var\left(\left(m_{full} - E\left[m_{full}|Z,X\right]\right)|D,Z,X\right)\right] + Var\left(E\left[m_{full}|D,Z,X\right]\right) - \Delta$$

where

$$\begin{aligned}
\Delta = & Var\left(\left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right)\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)\right) \\
& + Var\left((1-p_d)(\frac{R^D R^Y}{p_{11}} - 1)\left(E\left[Y|D,Z,X\right] - E\left[Y|Z,X\right]\right)\right) \\
& + E\left[\left(\frac{1-p_d^2}{p_{11}} - \frac{(1-p_d)^2}{p_{01}}\right) Var\left(Y|D,Z,X\right)\right] \\
& + 2E\left[(1-p_d) Cov\left(E\left[m_{full}|D,Z,X\right], E\left[Y|D,Z,X\right]|Z,X\right)\right]
\end{aligned}$$

Let n denote the sample size. Under certain regularity conditions, $\sqrt{n}$ convergence can be maintained with a non-parametric estimation in the first stage. We assume the well-established regularity conditions for Sieve basis functions that are presented in the literature (Newey (1994); Chen et al. (2008); Cattaneo (2010); Chaudhuri and Guilkey (2016)), and we construct $\sqrt{n}$ normality.

Theorem 1.5.5. *Let $\hat{E}(w)$ denote a vector of sieve estimation of first-stage nuisance estimators. For each component $e \in E$, suppose e is a function of*

d_e elements and is s_e times differentiable. Let $\eta = 1$ for power series basis; and $\eta = \frac{1}{2}$ for spline basis. Let K denote the terms in the series estimator, n denote the sample size, and $K = n^\nu$ such that:

$$4\eta + 2 < \frac{1}{\nu} < 4\frac{s_e}{d_e} - 6\eta$$

then,

(1) Under Assumption MAR, Overlap, and M,

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, (G'WG)^{-1}G'WV_{MAR}WG(G'WG)^{-1})$$

Furthermore, if $W = V_{MAR}^{-1}$,

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, (G'V_{MAR}^{-1}G)^{-1})$$

(2) Under Assumption SMAR, Overlap, and M,

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, (G'WG)^{-1}G'WV_{SMAR}WG(G'WG)^{-1})$$

Furthermore, if $W = V_{SMAR}^{-1}$,

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, (G'V_{SMAR}^{-1}G)^{-1})$$

The regularity conditions used here are no different from the conditions used in the work of Cattaneo (2010), Chaudhuri and Guilkey (2016). These conditions restrict the estimation of nuisance parameters to converge fast enough that they do not affect the convergence rate of the second-step estimation.

1.5.2 Efficiency

The well-known efficiency property of the AIPW estimator usually fails with the non-monotone missing pattern (Tsiatis (2007); Chaudhuri and Guilkey (2016)). We show two sufficient conditions in which the estimator maintains the efficiency property.

Lemma 1. *Suppose Assumption SMAR and Overlap hold, and* $Y \perp\!\!\!\perp D|Z, X$,

$$V_{SMAR} = V_{MAR}$$

Theorem 1.5.6. *Suppose Assumption Overlap, Assumption M, and one of the following assumptions hold:*

(i) Assumption MAR

(ii) Assumption SMAR and $Y \perp\!\!\!\perp D|Z, X$

then for β^0*, the asymptotic variance lower bound for* $\sqrt{N}(\hat{\beta} - \beta^0)$ *of any regular estimator* $\hat{\beta}$ *is given by* $\Omega = (G' V_{MAR}^{-1} G)^{-1}$*. An estimator with*

an asymptotic variance that is equal to Ω *has the following asymptotic linear representation:*

$$\sqrt{N}(\hat{\beta} - \beta^0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i, X_i, R_i^D, R_i^Y, R_i^D D_i, R_i^Y Y_i)$$

where

$$\psi(Z, X, R^D, R^Y, R^D D, R^Y Y) = -\Omega^{-1} G' V_{MAR}^{-1} m_{aipw}(Z, X, R^D, R^Y, R^D D, R^Y Y; \beta^0)$$

Although there are two sufficient conditions under which semiparametric efficiency bound holds, condition (ii) only holds when D does not affect the Y conditional of (Z, X) and is a strong assumption. In the settings with different missing variables, however, (ii) can be more reasonable. For example, when the fully observed variables are D and X, and the partially missing variables are Z and Y, and it is reasonable to assume that $Z \perp\!\!\!\perp Y|D, X$.

1.6 Discussion

The previous sections focused on missing treatment and outcome variables with a strict non-monotone missing pattern; this method can be generated to a more general arrangement. This section will discuss ways to extend the current approach.

1.6.1 Generalization to General Missing Patterns

The previous sections only considered the case where the missing pattern is strictly non-monotone and can be extended to a more general case that allows for monotone missingness, as well as univariate missingness. First, we propose a weaker overlap assumption.

Assumption Weak Overlap. $p_{11} \in [\kappa_{11}, 1)$ *almost surely in* (D, Z, X).

We provide a new AIPW moment function denoted by $\tilde{m}$ such that:

$$\tilde{m}_{aipw} = \frac{R^D R^Y}{p_{11}} Z(Y - g(D, X; \beta)) + \tilde{\phi}(Z, X, R^D, R^Y, R^D D, R^Y Y, \beta) \quad (1.6.1)$$

where

$$\begin{aligned}\tilde{\phi} = {} & p_{01}\left(\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_{11}}\right) Z\left[(Y - E\left[g(D, X; \beta)|Z, X\right]) - (E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right])\right] \\ & + \left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right) Z\left[(E\left[Y|D, Z, X\right] - g(D, X; \beta)) - (E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right])\right] \\ & + \left(1 - \frac{R^D R^Y}{p_{11}}\right) Z\left(E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right]\right) \qquad (1.6.2)\end{aligned}$$

Theorem 1.6.1. *Suppose Assumption MAR or SMAR, and Assumption Weak Overlap hold,* $E\left[\tilde{m}_{aipw}\right] = 0$ *when* $\beta = \beta^0$.

Proof of Theorem 1.6.1 follows the same argument as in the proof of Theorem 1.4.1.

For the previous moment condition to hold, we require p_d to be strictly positive in Z, X; and p_y to be strictly positive in Z, X when $R^D = 0$ and in Z, X, D when $R^D = 1$. Therefore, we need:

$$p_{01} = \Pr\left[R^D = 0, R^Y = 1 | Z, X\right] > 0$$
$$p_{11} = \Pr\left[R^D = 1, R^Y = 1 | Z, X, D\right] > 0$$

To allow for the case where $p_{01} = 0$, we multiply the augmentation term involving p_y by p_{01}. Adding in p_{01} does not affect the consistency of the estimator; thus, the first-stage estimation thereof does not affect the efficiency of the primary model.

1.6.2 Generalization to More than Two Missing Variables

This method can be extended to a situation in which there exists additional missing variables. We take missing IV, missing treatment status, and missing outcome as an example to illustrate the idea of extending the current method to more than two missing values. We use R^Z as the indicator of observing Z:

$$R^Z = \begin{cases} 1 & Z \text{ is observed} \\ 0 & Z \text{ is unobserved} \end{cases}$$

For simplicity, we only develop the moment condition under the SMAR-like assumption. Suppose there exists a vector of fully observed variables X,

and we make two analogous ignorability assumptions on the missing mechanism, and the overlap assumption is:

Assumption 1.6.1.

$$R^Z \perp\!\!\!\perp (D, Y)|X$$
$$R^D \perp\!\!\!\perp (D, Y)|X, R^Z Z, R^Z$$
$$R^Y \perp\!\!\!\perp (D, Y)|X, R^Z Z, R^D D, R^Z, R^D$$

Assumption 1.6.2. *(a)* $p_z \in [\kappa_z, 1)$ *almost surely in* (X)

(b) $p_d \in [\kappa_d, 1)$ *almost surely in* $(X, R^Z Z, R^Z)$

(c) $p_y \in [\kappa_y, 1)$ *almostly surely in* $(X, R^Z Z, R^D D, R^Z, R^D)$ *for* $\kappa_z > 0, \kappa_d > 0$,$\kappa_y > 0$.

The moment function is correspondingly:

$$m_{aipw} = \frac{R^Z R^D R^Y}{p_{111}} Z\left(Y - g(D, X; \beta)\right) + \phi(Z, X, R^Z, R^D, R^Y, R^Z Z, R^D D, R^Y Y, \beta)$$

where

$$
\begin{aligned}
\phi = & \left(1 - \frac{R^Z}{p_z}\right) \frac{R^D}{p_d} \frac{R^Y}{p_y} \left[E\left[Z|X\right]\left(Y - g(D, X; \beta)\right) - E\left[m_{full}|X\right]\right] \\
& + \frac{R^Z}{p_z} \left(1 - \frac{R^D}{p_d}\right) \frac{R^Y}{p_y} \left[Z\left(Y - E\left[g(D, X; \beta)|Z, X\right]\right) - E\left[m_{full}|X\right]\right] \\
& + \frac{R^Z}{p_z} \frac{R^D}{p_d} \left(1 - \frac{R^Y}{p_y}\right) \left[Z\left(E\left[Y|D, Z, X\right] - g(D, X; \beta)\right) - E\left[m_{full}|X\right]\right] \\
& + \left(1 - \frac{R^Z}{p_z}\right) \left(1 - \frac{R^D}{p_d}\right) \frac{R^Y}{p_y} \left[E\left[Z|X\right]\left(Y - E\left[g(D, X; \beta)|X\right]\right) - E\left[m_{full}|X\right]\right] \\
& + \left(1 - \frac{R^Z}{p_z}\right) \frac{R^D}{p_d} \left(1 - \frac{R^Y}{p_y}\right) \left[E\left[Z|X\right]\left(E\left[Y|D, X\right] - g(D, X; \beta)\right) - E\left[m_{full}|X\right]\right] \\
& + \frac{R^Z}{p_z} \left(1 - \frac{R^D}{p_d}\right) \left(1 - \frac{R^Y}{p_y}\right) \left[Z\left(E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right]\right) - E\left[m_{full}|X\right]\right] \\
& + \left(1 - \frac{R^Z R^D R^Y}{p_{111}}\right) E\left[m_{full}|X\right]
\end{aligned}
$$

where

$$
p_{111} = \Pr\left[R^Z = 1, R^D = 1, R^Y = 1|X, Z, D\right]
$$

In the case where there are two partially missing variables, the moment function can be derived through a three-step AIPW imputation by imputing the functions of Z using the AIPW strategy after the second step.

Theorem 1.6.2. *Suppose Assumptions 1.6.1 and 1.6.2 hold,* $E\left[m_{aipw}\right] = 0$ *when* $\beta = \beta^0$.

The proof is analogous to the proof of Theorem 1.4.1.

1.7 Monte Carlo Simulation

The previous results suggest that the AIPW-GMM approach yields consistent and more efficient results.. This section provides numerical evidence of these properties.

We consider the full model:

$$Y_i = \alpha D_i + \beta X_i + \epsilon_i \equiv 0.3D_i + 0.5X_i + \epsilon_i$$
$$D_i = 1(0.1 + 0.3Z_i + 0.1X_i \geq u_i)$$

where D_i is a single-value endogenous variable; X_i is an exogenous variable; Z_i is an instrument variable; and ϵ_i and u_i are jointly normally distributed with $\gamma = corr(\epsilon_i, u_i)$. R^D is determined by fully observed Z, X; and R^Y is determined by Z, X and partially observed D. These variables are determined via the binary model stated below:

$$R^D = 1(p_d \geq u_{rd})$$
$$R^Y = 1(p_y \geq u_{ry})$$

where

$$p_d = 0.2 + 0.2X + 0.3Z$$
$$p_y = 0.3 - 0.05X + 0.2Z + 0.3R_D D$$

and u_{rd}, u_{ry} are correlated. Since R^Y is determined by the endogenous variable D, and is conditional on Z, X; u_{ry} is correlated with ϵ, which is allowed in the SMAR assumption.

Table 1.1 shows the simulation results with different values for the correlation between ϵ and u. Because D affects R^Y, and D is correlated with ϵ when it is an endogenous treatment variable, this implies that R^Y is also endogenous in the sense that R^Y is correlated with ϵ. The higher the correlation between ϵ and u is, the more endogenous R^Y is; the endogeneity of R^Y will affect the consistency of the estimator derived from the complete case analysis because the moment condition no longer holds. As is shown in the Table 1.1, the complete case estimator for α is more biased when $corr(\epsilon, u)$ is higher, while the other two estimators remain consistent. The other finding is that in all exercises, the AIPW estimators have smaller RMSE compared to the other estimation strategies. The IPW estimators have higher RMSE than the CC estimator in some cases, because they not only drop all the observations with incomplete data, but also run a nonparametric estimation with the limited data on the nuisance parameters.

Table 1.2 shows the simulation results when the imputed values $\hat{E}[Y|Z, X]$, $\hat{E}[Y|D, Z, X]$, or $\hat{E}[D|Z, X]$ are misspecified. The AIPW-GMM estimator was shown to be robust when the missing mechanisms were correctly specified in Section 1.5, and misspecified imputed values should not affect the performance of the estimator. We can observe some evidence of the theoretical result on robustness in Table 1.2 of the theoretical result on robustness. The final exer-

Table 1.1Monte Carlo Simulation with Different Values for $corr(\epsilon, u)$

	$\alpha = 0.3$			$\beta = 0.5$		
	$\hat{\alpha}$	Mean Bias	RMSE	$\hat{\beta}$	Mean Bias	RMSE
N = 1000, R = 500, $corr(\epsilon, u) = 0.8$						
Complete Case	0.1842	0.1158	0.1537	0.4973	0.0027	0.1180
IPW	0.3223	0.0223	0.1255	0.5017	0.0017	0.1901
AIPW	0.3246	0.0246	0.0878	0.4798	0.0202	0.0863
N = 1000, R = 500, $corr(\epsilon, u) = 0.5$						
Complete Case	0.2178	0.0822	0.1405	0.4949	0.0051	0.1252
IPW	0.2974	0.0026	0.1285	0.5050	0.0050	0.1803
AIPW	0.2993	0.0007	0.0932	0.4903	0.0097	0.1144
N = 1000, R = 500, $corr(\epsilon, u) = 0.3$						
Complete Case	0.2552	0.0448	0.1162	0.4992	0.0008	0.1293
IPW	0.3136	0.0136	0.1498	0.5405	0.0405	0.3948
AIPW	0.3028	0.0028	0.0915	0.4792	0.0208	0.0919

Table 1.2Monte Carlo Simulation with Misspecified Imputed Values

	$\alpha = 0.3$			$\beta = 0.5$		
	$\hat{\alpha}$	Mean Bias	RMSE	$\hat{\beta}$	Mean Bias	RMSE
N = 1000, R = 500, misspecified $E[Y\|Z, X], E[Y\|D, Z, X]$						
Complete Case	0.1840	0.1160	0.1523	0.4904	0.0096	0.1253
IPW	0.3277	0.0277	0.1491	0.5344	0.0344	0.1039
AIPW	0.3067	0.0067	0.0857	0.4812	0.0188	0.0867
N = 1000, R = 500, misspecified $E[D\|Z, X]$						
Complete Case	0.2178	0.0822	0.1405	0.4949	0.0051	0.1252
IPW	0.2997	0.0003	0.1295	0.5076	0.0076	0.1714
AIPW	0.2990	0.0010	0.0923	0.4925	0.0075	0.1145
N = 1000, R = 500, misspecified p_y						
Complete Case	0.1716	0.1284	0.1665	0.4914	0.0086	0.1206
IPW	0.1705	0.1295	0.1674	0.5035	0.0035	0.1416
AIPW	0.3649	0.0649	0.1166	0.4716	0.0284	0.1324

cise illustrates how the AIPW estimator can be biased when $\hat{p}_y$ is misspecified and is captured by a function of Z, X without the partially observed D; this is the result when the missing mechanism is wrongly specified, and also when the correlation between the treatment status and the missing mechanism of the outcome is ignored.

1.8 Application

The Oregon Health Insurance Experiment is a large scale social experiment, for which most of the data were collected via surveys. In 2008, a group of low-income individuals was randomly selected for the opportunity to apply for the Oregon Health Plan (OHP) Standard, which is a Medicaid-extension program to cover low-income adults who are not eligible for the OHP Plus, which covers children, pregnant women, and families enrolled in the Temporary Assistance to Needy Families Program. The OHP standard program was not open for applicants until 2008. Participants registered for the lottery and were randomly assigned to win, conditional on the number of household members on the waiting list. The lottery winners were asked to return their application form. Only 60.82% of the lottery winners chose to return their application forms, and only some of those applications were approved. As such, this produced an endogenous non-compliance problem.

The data sets used here were composed of four parts. The descriptive data set recorded lottery participants' basic information and administrative data on the lotteries. The researchers conducted three follow-up surveys to

collect information on health insurance, healthcare needs, experiments, and costs. The initial follow-up survey was conducted right after the experiment during the period of June to November 2008 and included 58,405 survey participants. The intermediate survey was conducted six months after the experiment for a subsample of initial survey participants and included 11,756 participants. The final survey was conducted a year after the experiment for the same group of people who participated in the initial survey. These three surveys were referred to as the 0m, 6m, and 12m surveys, respectively.

We select the variables from the descriptive data, the 0m survey data, and the 12m survey data. The summary statistics of the variables used in this example are showed in Table 1.3. In the experiment, 50.66% of the survey participants were randomly selected to win the lottery, and selection into the lottery was used as the instrument variable, which is random conditional on the number of people in a household. For the number-of-household- members variable, 1 represented a household with a single member, while 2 and 3 represent households with two and more than two members. The age varies from 20 to 63; therefore, the influence of Medicare is excluded. The genders are balanced in the experiment; approximately 55% of the lottery participants are female. Most of the respondents are from the metropolitan statistical area, and less than 10% of the participants required a non-English questionnaire.

The second block in the table records partially observed variables. There are 7,611 participants with observed treatment statuses. We choose enrollment into the OHP program, including both the OHP standard and the

OHP plus, as the treatment variable.[15]

There are three outcomes from the final stage survey, including out-of-pocket costs for medical care, the number of days when physical health was not good, and how physical health had changed in the past six months. We refer to the health change as "Worse Health" in the later regression table, because for this variable, the higher the value, the worse the health status has become.

1.8.1 Missing Pattern

We first show the missing pattern of the chosen treatment and outcome variables to confirm the non-monotone missing pattern; the main reason for missingness in this data set is non-response to the surveys. The non-response rates for the initial (0m) and final (12m) surveys are shown in Figure 1.3.

[15]We did not choose the recorded enrollment for the OHP Standard in the administrative data as the treatment variable for the following reasons. First, the application's approval took a long time, it took 277 days for some people to get their application forms approved, and this was just before the final round of the survey; therefore, the administrative data do not show precise enrollment statuses during the first round of the survey. Moreover, even though some people were granted late approval, the effective date for the insurance card did not change; the late receivers needed to renew their insurance card to get coverage, which involved another endogenous self-selection problem. This directly results in the fact that even though they were approved and notified before the final round survey, approximately 1,400 participants chose the "Not Covered" option in the final round of the survey, and approximately 160 participants did not know their exact status OHP insurance, even though they were already notified about the application decision.

A previous study (Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P Newhouse, Heidi Allen, Katherine Baicker and Oregon Health Study Group (2012)) used external administrative data to recover the treatment status and avoid the missing problem. Since our goal is to show how the survey data can be used to get a consistent estimate, we choose to use the variables recorded in the survey, instead of in other data sets.

Table 1.3Summary Statistics in the OHIE Data

	Count	Mean	Sd	Min	Max
Z: Selected in the lottery	58405	0.5066	0.5000	0	1
X: Number of people in household	58405	1.2926	.4608	1	3
X: Age	58405	39.8863	12.1226	20	63
X: Female	58405	0.5464	0.4978	0	1
X: Zip code in a metropolitan statistical area	58405	0.7658	0.4235	0	1
X: Individual requested English-language materials	58405	.9072	.2902	0	1
D: Currently have OHP insurance	7611	0.1566	0.3634	0	1
Y: Total out-of-pocket costs for medical care, last 6 months	22539	230.5913	539.7317	0	4740
Y: Number of days (out of the past 30) when physical health not good	21415	9.5070	10.8559	0	30
Y: How has your health changed: past 6 months	23443	2.1503	.6058	1	3
Observations	58405				

For both the 0m and 12m surveys, the response rates were less than 50%. Furthermore, 16.88% of participants responded to the 0m but not to the 12m survey, so it is highly possible to observe their treatment statuses, but not their outcome statuses; for the group that responded to the 12m but not the 0m survey, it is possible to observe the opposite.

Another important source of missingness is non-response to survey questions among responders. We choose the survey participants who returned both the 0m and the 12m surveys, and the missing rate of answers for the questions on treatment and outcome status are shown in Figure 1.4. 16,566 participants returned both stage surveys. However, we could not confirm the treatment status for 8.73% of the respondents after correction of the variable.[16]For the outcome variables, the non-response rates vary from 1.35% to 9.53% among survey responders. The question related to subjective healthiness (worse health) suffers the least from missingness, while the question that asks for clear memories of the exact number of days when physical health is not good has the highest percentage of missing.

The non-response to survey and survey questions result in non-monotone missing patterns, and these are shown in Figure 1.5. We observe four missing patterns for all three outcomes, and this is consistent with the strict non-monotone missingness.

[16]If the participants mentioned that they had been successfully enrolled in the OHP Standard, and the administrative data showed consistent enrollment status, the OHP enrollment variable is corrected to value 1, whether it was missing or not.

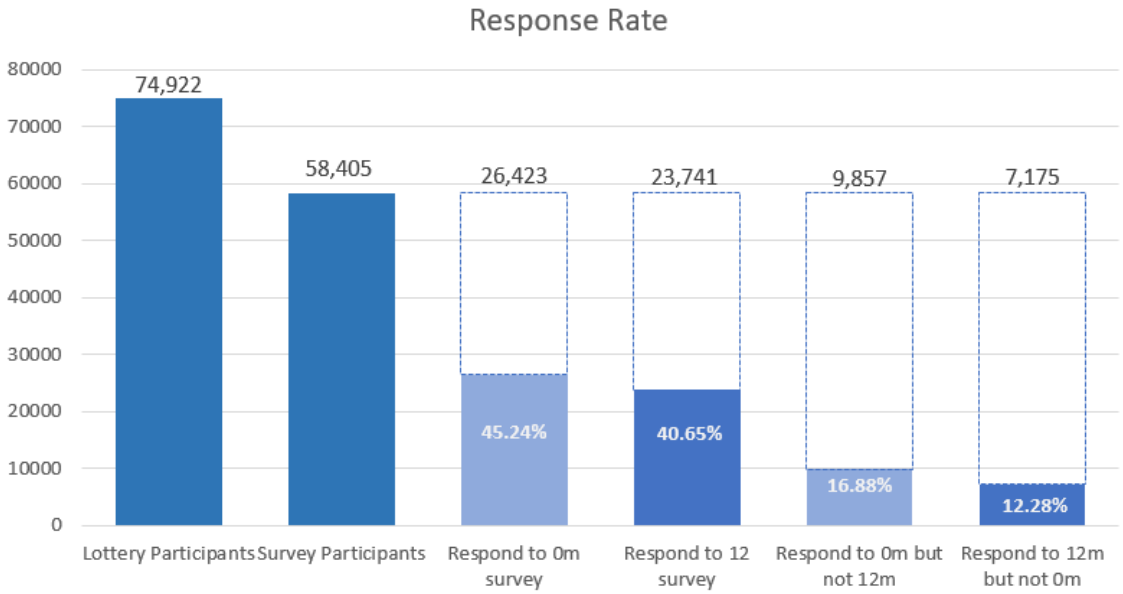


Figure 1.3. Non-response to the surveys

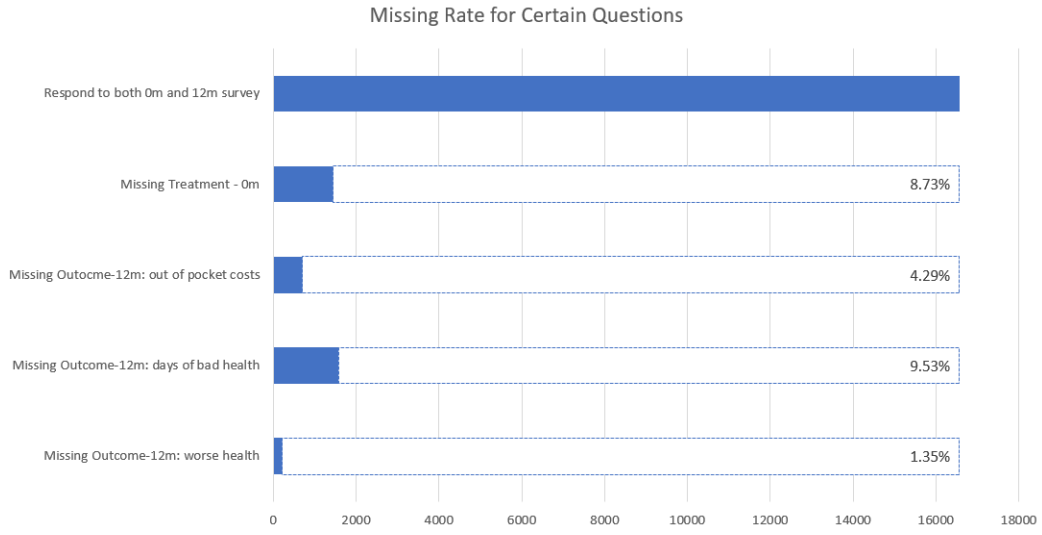


Figure 1.4. Non-response to survey questions

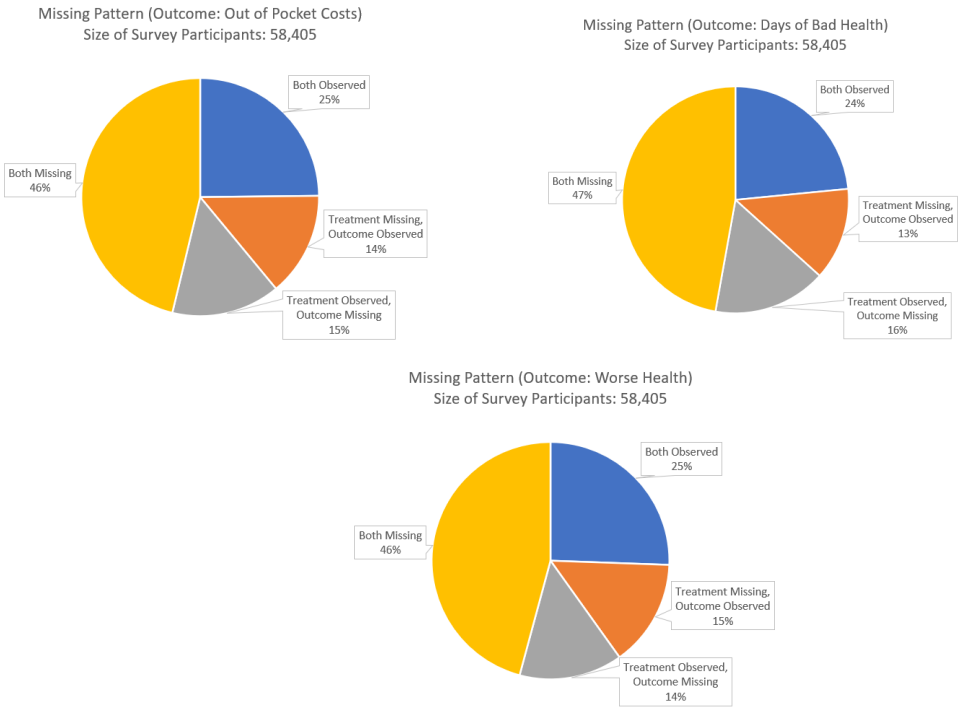


Figure 1.5. Non-monotone Missing Pattern

Next, we show evidence of a violation of the MAR assumption in this example. The MAR assumption assumes away the correlation between missingness in the outcomes and the partially observed treatment variable. We run a regression as a simple test on the correlation between them. As is shown in 1.4, when $R^D = 1$ (i.e., enrollment status in the OHP program is observed), there exists a significant correlation between R^Y and D. To use the SMAR assumption, it is also necessary to confirm that R^D does not depend on partially observe Y, so we run the same regression for the outcomes, and we find a small and insignificant correlation between R^D and Y when $R^Y = 1$; the results are recorded in Table 1.5.

1.8.2 Regression Results

Table 1.6 shows the regression results. The nuisance estimators are estimated by the sieve with B-spline basis functions, and the number of knots was selected through cross-validation. The estimated effects of OHP are shown in the first row. For the out-of-pocket costs, the CC estimator gives a lower number than the other two estimation strategies, while the IPW and AIPW estimators are closer to each other. The same pattern happens for the outcome of worse health, and the confidence interval of the CC estimator does not include the estimated values from IPW and AIPW approaches. This is evidence that the CC estimator can be biased in some circumstances; for the two outcomes introduced above, the CC estimator tends to overestimate the effect.

Table 1.4Regression of R^Y on D when $R^D = 1$

	Out-of-Pocket Costs	Days of Bad Health	Worse Health
Currently have OHP insurance	-0.0382***	-0.0393***	-0.0379***
	(0.00865)	(0.00880)	(0.00854)
Selected in the lottery	-0.0197**	-0.0124	-0.0207**
	(0.00657)	(0.00668)	(0.00648)
Number of people in household	0.0107	0.0150*	0.00953
	(0.00706)	(0.00718)	(0.00697)
Female	0.0419***	0.0391***	0.0437***
	(0.00644)	(0.00655)	(0.00636)
Age	0.00451***	0.00422***	0.00501***
	(0.000261)	(0.000265)	(0.000257)
Zip code in a metropolitan statistical area	0.00360	0.000263	-0.00327
	(0.00729)	(0.00741)	(0.00719)
Individual requested English-language materials	-0.0123	0.0370**	-0.0242
	(0.0127)	(0.0129)	(0.0125)
Constant	0.424***	0.350***	0.438***
	(0.0219)	(0.0222)	(0.0216)
Observations	23140	23140	23140

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.5Regression of R^D on Y when $R^Y = 1$

	Out-of-Pocket Costs	Days of Bad Health	Worse Health
Selected in the lottery	-0.00729	-0.00602	-0.00857
	(-1.15)	(-0.92)	(-1.37)
Number of people in household	0.00168	-0.00226	0.00365
	(0.24)	(-0.31)	(0.53)
Female	0.0464***	0.0495***	0.0454***
	(7.20)	(7.46)	(7.14)
Age	0.00420***	0.00421***	0.00414***
	(16.08)	(15.41)	(15.98)
Zip code in a metropolitan statistical area	-0.00584	-0.00500	-0.00734
	(-0.80)	(-0.66)	(-1.02)
Individual requested English-language materials	0.0951***	0.0898***	0.0998***
	(8.05)	(7.12)	(8.61)
Total out of pocket costs for medical care, last 6 months	1.64e-08		
	(0.81)		
Number of days (out of the past 30) when physical health not good		-0.000401	
		(-1.31)	
How has your health changed: past 6 months			-0.00820
			(-1.57)
Constant	0.351***	0.363***	0.366***
	(16.60)	(16.45)	(16.00)
Observations	22766	21415	23443

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We also find that AIPW estimators have smaller standard errors for all coefficient estimates than the other two estimation strategies; this is because the AIPW estimator tends to be significant at a higher significance level when the estimates are the same. Using the estimated coefficients on the OHP enrollment as an example, the standard errors were improved by 6–24% for different outcomes, compared to the CC and IPW estimators. For the days-of-bad-health outcome, the AIPW estimator is significant with a 99.9% significance level, while the other two estimators are significant at a lower significance level. However, this result was not apparent in the primary table, because the sample size was still as large as 14,500 when all observations with incomplete information are dropped. Therefore, we present another example in which the data are restricted to people who were 35–40 years of age, the days-of-bad-health outcome is shown in the table 1.7. When the sample size is small, the differences between the three estimates are more significant. The AIPW estimator is the only significant estimator with a significance level being 99%. The AIPW estimator gives an estimate lower than the other two estimation methods, while the IPW is closer to AIPW, compared to the CC estimator.

The results above show that enrolling in the OHP program reduced out-of-pocket costs by $199.7, reduced the number of days when physical health was not good by 3.3 days, and improved the health index by an average of 0.31. For survey participants who were 35–40 years of age, the effect of OHP enrollment on reducing days of bad physical health was greater than the overall

population, and on average, OHP enrollees tended to have five fewer days when their physical health was not good.

1.9 Conclusion

This paper studies the problem of two missing variables, and we focus on the missing treatment and outcome as an example, and we include a discussion of how to extend the current framework to more than two missing variables in Section 1.6. The first thing to do is to find the appropriate assumptions on the missing mechanism so that it can be identified. We propose the MAR assumption and the SMAR assumption, the difference between which lies in whether the correlation between the missing mechanism and partially observed variables is allowed; the identified missing mechanism is used in constructing an AIPW-GMM estimator.

Even though there are many desirable asymptotic properties recorded in the literature for the AIPW approach, many of these fail with non-monotone missingness. We find that these properties are maintained under the MAR assumption; they only hold under the SMAR assumption when the treatment variable has no direct effect on the outcome. The Monte Carlo simulation shows that the AIPW-GMM estimator performs better than the previously used CC and IPW estimators, in the sense that it is consistent and has the smallest standard error compared to the other two approaches. The performance is also verified in the empirical example of estimating the treatment effect of OHP on health-related outcomes. The AIPW-GMM estimator reduced

Table 1.6Regression Results

	Out-of-Pocket Costs			Days of Bad Health			Worse Health		
	(1) CC GMM	(2) IPW GMM	(3) AIPW GMM	(1) CC GMM	(2) IPW GMM	(3) AIPW GMM	(1) CC GMM	(2) IPW GMM	(3) AIPW GMM
OHP	-225.3***	-191.0***	-199.7***	-3.059**	-2.626**	-3.295***	-0.380***	-0.312***	-0.310***
	(38.59)	(35.11)	(32.99)	(1.053)	(0.890)	(0.805)	(0.0579)	(0.0508)	(0.0456)
Female	59.73***	46.96***	49.45***	-0.0935	0.199	0.140	-0.00485	0.00492	0.00734
	(6.963)	(10.75)	(6.044)	(0.191)	(0.193)	(0.149)	(0.0103)	(0.0107)	(0.00841)
Number of Household Members	-2.252	6.856	-5.612	-1.434***	-0.834***	-1.390***	-0.0418***	0.00311	-0.0351***
	(7.596)	(8.566)	(6.665)	(0.201)	(0.235)	(0.157)	(0.0108)	(0.0126)	(0.00900)
Age	1.079***	0.196	1.284***	0.164***	0.195***	0.168***	0.00457***	0.00629***	0.00558***
	(0.290)	(0.730)	(0.254)	(0.00750)	(0.00813)	(0.00609)	(0.000421)	(0.000453)	(0.000354)
MSA	-16.64*	-1.405	-26.33***	-0.972***	-0.657**	-0.957***	-0.0376**	-0.0102	-0.0225*
	(7.803)	(7.431)	(7.089)	(0.218)	(0.240)	(0.177)	(0.0115)	(0.0126)	(0.00981)
English-Speaking	32.60*	39.81*	21.17*	1.646***	2.427***	1.400***	0.159***	0.216***	0.186***
	(13.65)	(16.87)	(20.54)	(0.319)	(0.403)	(0.259)	(0.0171)	(0.0209)	(0.0149)
Constant	124.2***	68.29**	152.4***	4.112***	0.856	4.112***	1.950***	1.733***	1.850***
	(24.75)	(25.69)	(21.01)	(0.659)	(0.816)	(0.521)	(0.0357)	(0.0422)	(0.0306)
Observations	14500	14500	58396	13696	13696	58393	14933	14933	58358

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1.7Table: Effect of OHP on Days of Bad Health, for Age 35-40

	Days of Bad Health: Age 35-40		
	(1) CC GMM	(2) IPW GMM	(3) AIPW GMM
OHP	-3.305	-3.822	-5.855**
	(3.397)	(3.547)	(2.271)
Female	0.834	0.971	0.825
	(0.587)	(0.650)	(0.460)
Number of Household Members	-1.472*	-1.070	-1.624***
	(0.582)	(0.812)	(0.472)
Age	0.412*	1.333	0.472**
	(0.200)	(1.111)	(0.163)
MSA	-0.286	0.165	-1.270*
	(0.708)	(0.759)	(0.589)
English Speaking	0.106	0.589	-0.348
	(0.798)	(0.991)	(0.697)
Constant	-4.605	-40.06	-4.792
	(7.449)	(42.26)	(6.096)
Observations	1305	1305	6630

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

the standard error by 6-24% on the estimated coefficients for the treatment variables, compared to the CC and IPW estimators, and showed a significant effect of enrolling in the OHP on reducing out-of-pocket costs for medical care, reducing the number of days when physical health is not good, and improving health status.

Chapter 2

Sharp Bounds on Treatment Effects for Policy Evaluation

This is a joint work with Sukjin Han

2.1 Introduction

For counterfactual policy evaluation, it is important to ensure that treatment parameters are relevant to the policies in question. This is especially challenging in the presence of unobserved heterogeneity. This challenge is well featured in the definition of the local average treatment effect (LATE). The LATE has been one of the most popular treatment parameters used by empirical researchers since it was introduced by Guido W Imbens and Joshua D Angrist (1994). It induces a straightforward linear estimation method that requires only a binary instrumental variable (IV), and yet, allows for unrestricted treatment heterogeneity. The unfortunate feature of the LATE is that, as the name suggests, the parameter is intrinsically local, recovering the average treatment effect (ATE) for a specific subgroup of population called compliers. This feature leads to two major challenges in making the LATE a valuable parameter for counterfactual policy evaluation. First, the subpopulation for

which the effect is measured may not be the population of policy interest. Second, the definition of the subpopulation depends on the IV chosen, rendering the parameter even more difficult to extrapolate to new environments.

Dealing with the lack of external validity of the LATE has been an important theme in the literature. One approach in theoretical work (Joshua Angrist and Ivan Fernandez-Val (2010); Marinho Bertanha and Guido W Imbens (2019)) and empirical research (Rajeev Dehejia, Cristian Pop-Eleches and Cyrus Samii (2019); Karthik Muralidharan, Abhijeet Singh and Alejandro J Ganimian (2019)) has been to show the similarity between complier and non-complier groups based on observables. This approach, however, cannot attend to possible unobservable discrepancies between these groups. James J Heckman and Edward Vytlacil (2005) unify well-known treatment parameters by expressing them as weighted averages of what they define as the marginal treatment effect (MTE). This MTE framework has a great potential for extrapolation because a class of treatment parameters that are policy-relevant can also be generated as weighted averages of the MTE. The only obstacle is that the MTE is identified via a method called local IV (James J Heckman and Edward J Vytlacil (1999)), which requires the continuous variation of the IV that is sometime large depending on the targeted support. This in turn reflects the intrinsic difficulty of extrapolation when available exogenous variation is only discrete. Acknowledging this nature of the challenge, previous studies in the literature have proposed imposing shape restrictions on the MTE, which is a function of the treatment-selection unobservable, while

allowing for binary instruments in the framework of Heckman and Vytlacil (2005). Christian N Brinch, Magne Mogstad and Matthew Wiswall (2017*a*) introduce shape restrictions (e.g., linearity) on the MTE functions in an attempt to identify the LATE extrapolated to different subpopulations or to test for its externality validity. More recently, Magne Mogstad, Andres Santos and Alexander Torgovitsky (2018) propose a general partial identification framework where bounds on various policy-relevant treatment parameters can be obtained from a set of "IV-like estimands" that are directly identified from the data and routinely obtained in empirical work. Amanda E Kowalski (2020) applies an approach similar to these studies to extrapolate the results from one health insurance experiment to an external setting.

This paper continues this pursuit and investigates the possibility of extrapolating local treatment parameters to different policy settings in the MTE framework when IVs are only binary. In a partial identification framework similar in spirit to Mogstad, Santos and Torgovitsky (2018), we show how to systematically calculate sharp nonparametric bounds on various extrapolated treatment parameters for binary (or more generally, discrete) outcomes using instruments that are allowed to be binary. These parameters are defined as weighted averages of the MTE. Examples include the ATE, the treatment on the treated, the LATE for subgroups induced by new policies, and the policy-relevant treatment effect (PRTE). We also show how to place in this procedure restrictions from a large menu of identifying assumptions beyond the shape restrictions considered in earlier work.

In this paper, we make four main contributions. First, we propose a novel framework for calculating bounds on policy-relevant treatment parameters. We introduce the probability of the latent state of the outcome-generating process conditional on the treatment-selection unobservable. This latent conditional probability is the key ingredient for our analysis, as both the target parameter and the distribution of the observables can be written as linear functionals of it. Therefore, having it as a decision variable, we can formulate infinite-dimensional linear programming that produces bounds on a targeted treatment parameter. This approach is reminiscent of Alexander Balke and Judea Pearl (1997) and can be viewed as its generalization to the MTE framework. Balke and Pearl (1997) introduce a linear programming approach to characterize bounds on the ATE with a binary outcome, treatment and instrument. The main distinction of our approach is that the latent probability is conditioned on the selection unobservable, which is important for our extrapolation purpose. To make it feasible to solve the resulting infinite-dimensional program, we use a sieve-like approximation of the program and produce a finite-dimensional linear program (LP). This approximation approach builds on Mogstad, Santos and Torgovitsky (2018), although they use approximation directly on the MTE function. We also propose a conservative approach to choosing the sieve dimension in practice.

Second, by formulating the LP based on the latent conditional probability rather than the MTE, it creates a flexible environment where we can introduce identifying assumptions that have not been used in the context of

the MTE framework or the LATE extrapolation. We propose assumptions that there exist exogenous variables other than IVs. We propose two types of exogenous variables that have been used in the context of identifying the ATE in the literature: Ismael Mourifié (2015), Sukjin Han and Edward J Vytlacil (2017), Quang Vuong and Haiqing Xu (2017), and Sukjin Han and Sungwon Lee (2019) use the first type, and Edward Vytlacil and Nese Yildiz (2007), Azeem M Shaikh and Edward J Vytlacil (2011), and Balat and Han (2018) use the second type. We utilize these variables in this novel context of the MTE framework. Also, while the earlier papers exploit these variables in combination with rank similarity or rank invariance, we show that they independently have identifying power for treatment parameters, including the ATE. We also propose identifying assumptions such as uniformity and the direction of endogeneity in this MTE framework. The direction of endogeneity is sometimes assumed in empirical work to characterize selection bias and has been shown to have identifying power (Charles F Manski and John V Pepper (2000a)). The uniformity assumption is related to rank similarity or rank invariance (Victor Chernozhukov and Christian Hansen (2005)). The shape restrictions on the MTE considered in the literature can also be nested within our framework, since the MTE is just a sum of the latent conditional probabilities. The assumptions on the existence of exogenous variables complement the identifying assumptions that rely on the researcher's prior, in that its identifying power comes from actual data. When a confidence set is constructed under one of the latter assumptions, we can conduct a specification test for that assumption.

Third, we show that our approach yields straightforward proof of the sharpness of the resulting bounds. This feature stems from the use of the latent conditional probability in the linear programming and the convexity of the feasible set in the program. When the MTE itself is the target parameter, we distinguish between the notions of point-wise and uniform sharpness and argue why uniform sharpness is often difficult to achieve.

Fourth, as an application, we study the effects of insurance on medical service utilization by considering various counterfactual policies related to insurance coverage. The LATE for compliers and the bounds on the LATE for always-takers and never-takers reveal that possessing private insurance has the largest effect on medical visits for never takers, i.e., those who face higher insurance cost. This provides a policy implication that lowering the cost of private insurance is important, because the high cost might hinder people with most need from receiving adequate medical services.

The linear programming approach to partial identification of treatment effects was pioneered by Balke and Pearl (1997) and recently gained attention in the literature; see, e.g., Mogstad, Santos and Torgovitsky (2018), Alexander Torgovitsky (2019*a*), Cecilia Machado, Azeem Shaikh and Edward Vytlacil (2019), Vishal Kamat (2019), Florian Gunsilius (2019), and Sukjin Han (2020*b*).[1] As these papers suggest, there are many settings, including ours,

[1]For the computational approach in contexts other than program evaluation, see Charles F Manski (2007), Yuichi Kitamura and Jörg Stoye (2019), Rahul Deb, Yuichi Kitamura, John Kim-Ho Quah and Jörg Stoye (2017), and Pietro Tebaldi, Alexander Torgovitsky and Hanbin Yang (2019).

where analytical derivation of bounds is cumbersome or nearly impossible due to the complexity of the problems.

This paper will proceed as follows. The next section introduces the main observables, maintained assumptions, and target parameters. Section 2.3 defines the latent conditional probability and formulates the infinite-dimensional LP, and Section 2.4 introduces sieve approximation to the program. Section 2.5 then generalize the analysis to incorporate additional exogenous variables. Section 2.6 proposes a menu of identifying assumptions and shows how they can easily be incorporated in the LP. Section 2.7 provides numerical illustrations, and Section 2.8 contains an empirical application. In the Appendix, Section B.1 lists the examples of target parameters. Section B.2 discusses (i) the point-wise and uniform sharpness for the MTE bounds, (ii) inference, especially how to conduct specification tests for identifying assumptions, (iii) an extension with continuous covariates, and (iv) the relationship between this paper's LP and those in Mogstad, Santos and Torgovitsky (2018). All proofs are contained in Section B.3.

2.2 Observables and Target Parameters

Assume that we observe the binary outcome $Y \in \{0,1\}$, binary treatment $D \in \{0,1\}$, and binary instrument $Z \in \{0,1\}$. We may additionally observe (possibly endogenous) discrete covariates $X \in \mathcal{X}$.[2] Binary Y is com-

[2]We focus on discrete X as it simplifies the exposition. Section B.2.3 in the Appendix extends the framework to incorporate continuously distributed X.

mon in empirical work. Binary Z is also common, especially in randomized experiments, and allowing for this minimal exogenous variation is the key challenge for extrapolation that we want to address in this paper. Still, the analysis of this paper can be extended to allow for general discrete Y and Z. Let $Y(d)$ be the counterfactual outcome given $D = d$, which is consistent with the observed outcome: $Y = DY(1) + (1 - D)Y(0)$. We maintain the following assumptions:

Assumption SEL. $D = 1\{U \leq P(Z, X)\}$ *where* $P(Z, X) \equiv \Pr[D = 1|Z, X]$ *and* $U|_{X=x} \sim Unif[0, 1]$ *for* $x \in \mathcal{X}$.

Assumption EX. $(Y(d), D(z)) \perp Z|X$.

Assumption SEL imposes a selection model for D, which is important in motivating and interpreting marginal treatment effects later. This assumption is also equivalent to Imbens and Angrist (1994)'s monotonicity assumption (Edward Vytlacil (2002)). We introduce the standard normalization that $U \sim Unif[0, 1]$ conditional on $X = x$.[3] Assumption EX imposes the exclusion restriction and conditional independence for Z.

Heckman and Vytlacil (2005) establish that various treatment parameters can be expressed as integral equations of the MTE, defined as

$$E[Y(1) - Y(0)|U = u, X = x].$$

[3]Note that for any index function $g(z, x)$ and an unobservable ε with any distribution, the selection model satisfies $D = 1\{\varepsilon \leq g(Z, X)\} = 1\{F_{\varepsilon|X}(\varepsilon|X) \leq F_{\varepsilon|X}(g(Z, X)|X)\} = 1\{U \leq P(Z, X)\}$, since $P(z, x) = \Pr[\varepsilon \leq g(z, x)|X = x] = \Pr[U \leq F_{\varepsilon|X}(g(z, x)|x)|X = x] = F_{\varepsilon|X}(g(z, x)|x)$ and $F_{\varepsilon|X}(\varepsilon|X) = U$ is uniformly distributed conditional on X.

Following Mogstad, Santos and Torgovitsky (2018), it is convenient to introduce the marginal treatment response (MTR) function

$$\begin{aligned} m_d(u,x) &\equiv E[Y(d)|U=u,X=x] \\ &= \Pr[Y(d)=1|U=u,X=x]. \end{aligned}$$

Then, the MTE can be expressed as $m_1(u,x) - m_0(u,x)$. Now, we define the target parameter τ to be a weighted average of the MTE:

$$\tau = E[\tau_1(Z,X) - \tau_0(Z,X)], \tag{2.2.1}$$

where

$$\tau_d(z,x) = \int m_d(u,x) w_d(u,z,x) du \tag{2.2.2}$$

by using $F_{U|X}(u|x) = u$, and $w_d(u,z,x)$ is a known weight specific to the parameter of interest.[4] This definition agrees with the insight of Heckman and Vytlacil (2005). The target parameter includes a wide range of policy-relevant treatment parameters. With a Dirac delta function for a given value u as the weight, the MTE itself can be an example. We list a few examples of the target parameter here; other examples can be found in Table B.1 in the Appendix.

Example 3. *The ATE can be a target parameter with* $w_d(u,z,x) = 1, \forall u,z,x$.

$$\tau_{ATE} = E\left[\int_0^1 m_1(u,X)du - \int_0^1 m_0(u,X)du\right]$$

[4]Mogstad, Santos and Torgovitsky (2018) define the weight in a slightly different way.

Example 4. *The generalized LATE for always-takers and never-takers are also target parameters. Here, we give the expression of the LATE for always-takers as an example. Assume $P(z,x)$ increases in z for any given $x \in \mathcal{X}$. For the always-taker (AT) LATE, we give weight $\frac{1}{P(0,x)}$ to individuals with $u \in [0, P(0,x)]$ and thus, we have $w_d(u,z,x) = \frac{1(u \in [0,p(0,x)])}{p(0,x)}$.*

$$\tau_{LATE\text{-}AT} = E\left[\int_0^1 m_1(u,X)\frac{1(u \in [0,p(0,X)])}{p(0,X)}du - \int_0^1 m_0(u,X)\frac{1(u \in [0,p(0,X)])}{p(0,X)}du\right]$$

Example 5. *The policy relevant treatment effect (PRTE) is a target parameter that is particularly useful for policy evaluation. It is defined as the welfare difference between two different policies. Let Z and Z' be two instrument variables under two policies and $P(Z,X)$ and $P'(Z',X)$ be propensity scores under the two policies.*

$$\begin{aligned}\tau_{PRTE} = E\Bigg[&\int_0^1 m_1(u,X)\frac{\Pr\left[u \le P'(Z',X)\right] - \Pr\left[u \le P(Z,X)\right]}{E\left[P'(Z',X)\right] - E\left[P(Z,X)\right]}du \\ &- \int_0^1 m_0(u,X)\frac{\Pr\left[u \le P'(Z',X)\right] - \Pr\left[u \le P(Z,X)\right]}{E\left[P'(Z',X)\right] - E\left[P(Z,X)\right]}du\Bigg]\end{aligned}$$

In these examples, the weights w_0 and w_1 can be set asymmetrically to define a broader class of parameters. All the parameters we consider in this paper can be defined conditional on X, although we omit them for succinctness.

2.3 Distribution of Latent State and Infinite-Dimensional Linear Program

As a crucial first step of our analysis, we define a state variable that determines a specific mapping of

$$d \mapsto y.$$

Since $d \in \{0, 1\}$ and $y \in \{0, 1\}$, there are four possible maps from d onto y. Define a discrete latent variable ϵ whose value e corresponds to each possible map:

$$\epsilon \in \mathcal{E},$$

where $|\mathcal{E}| = 4$ with $\mathcal{E} \equiv \{1, 2, 3, 4\}$. That is, ϵ is a decimal transformation of a binary sequence $(Y(1), Y(0))$, which captures the treatment effect heterogeneity. For the later purpose, it is helpful to explicitly define the map as

$$y = g_e(d)$$

and write

$$Y(d) = g_\epsilon(d), \tag{2.3.1}$$

which implies $Y = g_\epsilon(D)$. It is important to note that no structure is imposed in introducing $g_e(\cdot)$ because the mapping is saturated by binary Y and D. By (2.3.1) and Assumption SEL, Assumption EX can be equivalently stated

as $(\epsilon, U) \perp Z|X$. Still, ϵ and X can be correlated as X is allowed to be endogenous.

Now, as a key component of our LP, we define the probability mass function of ϵ conditional on (U, X): for $e \in \mathcal{E}$,

$$q(e|u, x) \equiv \Pr[\epsilon = e|U = u, X = x] \tag{2.3.2}$$

with $\sum_{e \in \mathcal{E}} q(e|u, x) = 1$ for any u, x. The quantity $q(e|u, x)$ captures endogenous treatment selection. It is shown below that this latent conditional probability is a building block for various treatment parameters and thus serves as the decision variable in the LP. The introduction of $q(e|u, x)$ distinguishes our approach from those in Balke and Pearl (1997) and Mogstad, Santos and Torgovitsky (2018). Since the probability is conditional on continuously distributed U, the simple finite-dimensional linear programming approach of Balke and Pearl (1997) is no longer applicable. Instead, we use an approximation method similar to Mogstad, Santos and Torgovitsky (2018). However, Mogstad, Santos and Torgovitsky (2018) uses the MTR function as a building block for treatment parameters and introduces the "IV-like" estimands as a means of funneling the information from the data. Unlike in Mogstad, Santos and Torgovitsky (2018), $q(e|u, x)$ can be directly related to the distribution of data. This allows us to later incorporate identifying assumptions that are difficult to incorporate within the framework of Mogstad, Santos and Torgovitsky (2018).

By (2.3.1) and (2.3.2), note that

$$\begin{aligned}\Pr[Y(d)=1|U=u,X=x] &= \Pr[\epsilon \in \{e \in \mathcal{E} : g_e(d)=1\}|U=u,X=x] \\ &= \sum_{e\in\mathcal{E}:g_e(d)=1} q(e|u,x).\end{aligned}$$

Therefore, the MTR can be expressed as

$$m_d(u,x) = \sum_{e:g_e(d)=1} q(e|u,x). \tag{2.3.3}$$

Combining (2.3.3) and (2.2.2), we have $\tau_d(z,x) = \sum_{e:g_e(d)=1} \int q(e|u,x)w_d(u,z,x)du$, and thus the target parameter $\tau = E[\tau_1(Z,X)] - E[\tau_0(Z,X)]$ in (2.2.1) can be written as

$$\tau = \sum_{e:g_e(1)=1} \int E[q(e|u,X)w_1(u,Z,X)]du - \sum_{e:g_e(0)=1} \int E[q(e|u,X)w_0(u,Z,X)]du \tag{2.3.4}$$

for some q that satisfies the properties of probability.

The goal of this paper is to (at least partially) infer the target parameter τ based on the data, i.e., the distribution of (Y,D,Z,X). The key insight is that there are observationally equivalent $q(e|u,x)$'s that are consistent with the data, which in turn produces observationally equivalent τ's that define the identified set.

Let $p(y,d|z,x) \equiv \Pr[Y=y,D=d|Z=z,X=x]$ be the observed conditional probability. This data distribution imposes restrictions on $q(e|u,x)$. For instance, for $D=1$,

$$\begin{aligned}p(y,1|z,x) &= \Pr[Y(1)=y, U \leq P(z,x)|Z=z,X=x] \\ &= \Pr[Y(1)=y, U \leq P(z,x)|X=x]\end{aligned}$$

by Assumption EX, but

$$\Pr[Y(1)=y, U\leq P(z,x)|X=x] = \int_0^{P(z,x)} \Pr[Y(1)=y|U=u,X=x]du$$
$$= \sum_{e:g_e(1)=y} \int_0^{P(z,x)} q(e|u,x)du, \qquad (2.3.5)$$

where the second equality is by $\Pr[Y(d)=y|U=u,X=x] = \sum_{e:g_e(d)=y} q(e|u,x)$.

To define the identified set for τ, we introduce some simplifying notation. Let $q(u,x) \equiv \{q(e|u,x)\}_{e\in\mathcal{E}}$ and

$$\mathcal{Q} \equiv \{q(\cdot) : \sum_{e\in\mathcal{E}} q(e|u,x) = 1 \,\forall(u,x) \text{ and } q(e|u,x)\geq 0 \,\forall(e,u,x)\}$$

be the class of $q(u,x)$, and let $p \equiv \{p(1,d|z,x)\}_{(d,z,x)\in\{0,1\}^2\times\mathcal{X}}$. Also, let $R_\tau : \mathcal{Q}\to\mathbb{R}$ and $R_0 : \mathcal{Q}\to\mathbb{R}^{d_p}$ (with d_p being the dimension of p) denote the linear operators of $q(\cdot)$ that satisfy

$$R_\tau q \equiv \sum_{e:g_e(1)=1} \int E[q(e|u,X)w_1^\tau(u,Z,X)]du - \sum_{e:g_e(0)=1} \int E[q(e|u,X)w_0^\tau(u,Z,X)]du,$$
$$R_0 q \equiv \sum_{e:g_e(d)=1} \int_{\mathcal{U}_{z,x}^d} q(e|u,x)du,$$

where $\mathcal{U}_{z,x}^d$ denotes the intervals $\mathcal{U}_{z,x}^1 \equiv [0,P(z,x)]$ and $\mathcal{U}_{z,x}^0 \equiv (P(z,x),1]$.

Definition 2.3.1. *The identified set of τ is defined as*

$$\mathcal{T}^* \equiv \{\tau\in\mathbb{R} : \tau = R_\tau q \text{ for some } q\in\mathcal{Q} \text{ such that } R_0 q = p\}.$$

In what follows, we formulate the infinite-dimensional LP (∞-LP) that characterizes $\mathcal{T}^*$. This program conceptualizes sharp bounds on τ from the

data and the maintained assumptions (Assumptions SEL and EX). The upper and lower bounds on τ are defined as

$$\overline{\tau} = \sup_{q \in \mathcal{Q}} R_\tau q, \tag{∞-LP1}$$

$$\underline{\tau} = \inf_{q \in \mathcal{Q}} R_\tau q, \tag{∞-LP2}$$

subject to

$$R_0 q = p. \tag{∞-LP3}$$

Observe that the set of constraints (∞-LP3) does not include

$$\sum_{e: g_e(d)=0} \int_{\mathcal{U}^d_{z,x}} q(e|u,x)du = p(0,d|z,x) \qquad \forall (d,z,x) \in \{0,1\}^2 \times \mathcal{X}. \tag{2.3.6}$$

This is because we know a priori that they are redundant in the sense that they do not further restrict the *feasible set*, i.e., the set of $q(e|u,x)$'s that satisfy all the constraints ($q \in \mathcal{Q}$ and (∞-LP3)).

Lemma 2. *In the linear program* (∞-LP1)–(∞-LP3)*, the feasible set defined by* $q \in \mathcal{Q}$ *and* (∞-LP3) *is identical to the feasible set defined by* $q \in \mathcal{Q}$*,* (∞-LP3)*, and* (2.3.6)*.*

Theorem 2.3.1. *Under Assumptions SEL and EX, suppose* $\mathcal{T}^*$ *is non-empty. Then, the bounds* $[\underline{\tau}, \overline{\tau}]$ *in* (∞-LP1)–(∞-LP3) *are sharp for the target parameter* τ*, i.e.,* $cl(\mathcal{T}^*) = [\underline{\tau}, \overline{\tau}]$*, where* $cl(\cdot)$ *is the closure of a set.*

The result of this theorem is immediate due to the convexity of the feasible set $\{q : q \in \mathcal{Q}\} \cap \{q : R_0 q = p\}$ in the LP and the linearity of $R_\tau q$ in q, which implies that $[\underline{\tau}, \overline{\tau}]$ is convex.

2.4 Sieve Approximation and Finite-Dimensional Linear Programming

Although conceptually useful, the LP (∞-LP1)–(∞-LP3) is not feasible in practice because $\mathcal{Q}$ is an infinite-dimensional space. In this section, we approximate (∞-LP1)–(∞-LP3) with a finite-dimensional LP via a sieve approximation of the conditional probability $q(e|u,x)$. We use Bernstein polynomials as the sieve basis. Bernstein polynomials are useful in imposing restrictions on the original function (Kenneth I Joy (2000); Xiaohong Chen, Elie T Tamer and Alexander Torgovitsky (2011); Xiaoyan Chen, Jieqing Tan, Zhi Liu and Jin Xie (2017)) and therefore have been introduced in the context of linear programming (Mogstad, Santos and Torgovitsky (2018); Matthew A Masten and Alexandre Poirier (2018); Magne Mogstad, Alexander Torgovitsky and Christopher R Walters (2019)).

Consider the following sieve approximation of $q(e|u,x)$ using Bernstein polynomials of order K

$$q(e|u,x) \approx \sum_{k=1}^{K} \theta_k^{e,x} b_k(u),$$

where $b_k(u) \equiv \binom{K}{k} x^k (1-x)^{K-k}$ is a univariate Bernstein basis, $\theta_k^{e,x} \equiv \theta_{k,K}^{e,x} \equiv q(e|k/K,x)$ is its coefficient, and K is finite. It is important to note that x can index θ, because $q(e|u,x)$ is a saturated function of x. By the definition of the Bernstein coefficient, for any (e,x), it satisfies $q(e|u,x) \geq 0$ for all u if and only if $\theta_k^{e,x} \geq 0$ for all k. Also, $\sum_{e\in\mathcal{E}} q(e|u,x) = 1$ for all (u,x) is approximately equivalent to $\sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1$ for all (k,x). To see this, first,

$\sum_{e\in\mathcal{E}} q(e|u,x) = 1$ for all (u,x) implies $\sum_{e\in\mathcal{E}} \theta_k^{e,x} = \sum_{e\in\mathcal{E}} q(e|k/K,x) = 1$ for all (k,x). Conversely, when $\sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1$ for all (k,x),

$$\sum_{e\in\mathcal{E}} q(e|u,x) \approx \sum_{e\in\mathcal{E}} \sum_{k=1}^{K} \theta_k^{e,x} b_k(u) = \sum_{k=1}^{K} b_k(u) = 1$$

by the binomial theorem (Julian L Coolidge (1949)). Motivated by this approximation, we formally define the following sieve space for $\mathcal{Q}$:

$$\mathcal{Q}_K \equiv \left\{ \left\{ \sum_{k=1}^{K} \theta_k^{e,x} b_k(u) \right\}_{e\in\mathcal{E}} : \sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1 \,\forall (k,x) \text{ and } \theta_k^{e,x} \geq 0 \,\forall (e,k,x) \right\} \subseteq \mathcal{Q}. \tag{2.4.1}$$

Let $\mathcal{K} \equiv \{1,...,K\}$ and $p(z,x) \equiv \Pr[Z=z, X=x]$. For $q \in \mathcal{Q}_K$, by (2.3.4) and (2.4.1), the target parameter $\tau = E[\tau_1(Z,X)] - E[\tau_0(Z,X)]$ can be expressed with

$$\begin{aligned} E[\tau_d(Z,X)] &= \sum_{e:g_e(d)=1} \sum_{(k,x)\in\mathcal{K}\times\mathcal{X}} \theta_k^{e,x} \int b_k(u) \sum_{z\in\{0,1\}} w_d(u,z,x) p(z,x) du \\ &\equiv \sum_{e:g_e(d)=1} \sum_{(k,x)\in\mathcal{K}\times\mathcal{X}} \theta_k^{e,x} \gamma_k^d(x), \end{aligned} \tag{2.4.2}$$

where $\gamma_k^d(x) \equiv \int b_k(u) \sum_{z\in\{0,1\}} w_d(u,z,x) p(z,x) du$. Also, for $q \in \mathcal{Q}_K$ and $D = 1$, by (2.3.5), we have

$$\begin{aligned} p(y,1|z,x) &= \sum_{e:g_e(d)=y} \sum_{k\in\mathcal{K}} \theta_k^{e,x} \int_0^{P(z,x)} b_k(u) du \\ &\equiv \sum_{e:g_e(d)=y} \sum_{k\in\mathcal{K}} \theta_k^{e,x} \delta_k^1(z,x), \end{aligned} \tag{2.4.3}$$

where $\delta_k^d(z,x) \equiv \int_{\mathcal{U}_{z,x}^d} b_k(u) du$.

From (2.4.2) and (2.4.3), we can expect that a finite-dimensional LP can be obtained with respect to $\theta_k^{e,x}$. Let $\theta \equiv \{\theta_k^{e,x}\}_{(e,k,x)\in\mathcal{E}\times\mathcal{K}\times\mathcal{X}}$ and let

$$\Theta_K \equiv \left\{ \theta : \sum_{e\in\mathcal{E}} \theta_k^{e,x} = 1 \,\forall(k,x) \text{ and } \theta_k^{e,x} \geq 0 \,\forall(e,k,x) \right\}.$$

Then, we can formulate the following finite-dimensional LP that corresponds to the ∞-LP in (∞-LP1)–(∞-LP3):

$$\overline{\tau}_K = \max_{\theta\in\Theta_K} \sum_{(k,x)\in\mathcal{K}\times\mathcal{X}} \left\{ \sum_{e:g_e(1)=1} \theta_k^{e,x}\gamma_k^1(x) - \sum_{e:g_e(0)=1} \theta_k^{e,x}\gamma_k^0(x) \right\} \tag{LP1}$$

$$\underline{\tau}_K = \min_{\theta\in\Theta_K} \sum_{(k,x)\in\mathcal{K}\times\mathcal{X}} \left\{ \sum_{e:g_e(1)=1} \theta_k^{e,x}\gamma_k^1(x) - \sum_{e:g_e(0)=1} \theta_k^{e,x}\gamma_k^0(x) \right\} \tag{LP2}$$

subject to

$$\sum_{e:g_e(d)=1} \sum_{k\in\mathcal{K}} \theta_k^{e,x}\delta_k^d(z,x) = p(1,d|z,x) \qquad \forall(d,z,x) \in \{0,1\}^2 \times \mathcal{X}. \tag{LP3}$$

This LP is computationally very easy to solve using standard algorithms, such as the simplex algorithm; conditional on x, when $K = 50$ and $\dim(\theta) = 204$, it takes only around 10 seconds to calculate $\overline{\tau}_K$ and $\underline{\tau}_K$ with moderate computing power. The important remaining question is how to choose K in practice. We discuss this issue in Section 2.7. Finally, it is worth noting that, extending Proposition 4 in Mogstad, Santos and Torgovitsky (2018), we may exactly calculate $\overline{\tau}$ and $\underline{\tau}$ (i.e., $\overline{\tau} = \overline{\tau}_K$ and $\underline{\tau} = \underline{\tau}_K$) under the assumptions that (i) the weight function $w_d(u,z,x)$ is piece-wise constant in u and (ii) the constant spline that provides the best mean squared error approximation of $q(e|u,x)$ satisfies all the maintained assumptions (possibly including the identifying assumptions introduced later) that $q(e|u,x)$ itself satisfies; see Mogstad, Santos and Torgovitsky (2018) for details.

2.5 General Analysis

Now we generalize the analysis in Sections 2.2–2.4 to incorporate additional exogenous variables other than the instrument Z that researchers may be equipped with. We show that these variables are fruitful for narrowing bounds on the target parameter. This is the first paper that introduces this type of variable in the MTE framework. This is also the first paper that shows the usefulness of these variables without necessarily combining them with assumptions related to rank similarity or rank invariance.

Let $W \in \mathcal{W}$ be such an exogenous variable. We assume that W is discrete. We show that even binary variation in W can be useful in improving the bounds. We modify our maintained assumptions to consider two different scenarios related to W: (a) W directly affects Y but not D and (b) W directly affects both Y and D. Let $Y(d, w)$ be the extended counterfactual outcome of Y given (d, w).

Assumption SEL$_W$. *(a) Assumption SEL; (b)* $D = 1\{U \leq P(Z, X, W)\}$ *where* $P(Z, X, W) \equiv \Pr[D = 1|Z, X, W]$.

Assumption EX$_W$. *(a)* $(Y(d, w), D(z)) \perp (Z, W)|X$*; (b)* $(Y(d, w), D(z, w)) \perp (Z, W)|X$.

Case (a) is where W is a reversely excluded exogenous variable, which we call *reverse IV*. This type of exogenous variables was considered by Vytlacil and Yildiz (2007), Shaikh and Vytlacil (2011), and Balat and Han (2018). However, unlike those studies, we exploit W without rank similarity or rank

invariance. In Case (b), we show that a reverse IV is not necessary, and W can be present in the selection equation. This type of exogenous variables was considered by Mourifié (2015), Han and Vytlacil (2017), Vuong and Xu (2017), and Han and Lee (2019), but again, unlike these papers, we do not necessarily assume rank similarity or rank invariance. Below, we combine the existence of W (for both scenarios) with assumptions that are related to rank similarity. Another distinct feature of our approach in comparison to the prior studies is that we consider a broad class of the generalized LATEs as our target parameter, including the ATE considered in those studies.

In what follows, we modify the linear programming framework from Sections 2.2 and 2.3 to reflect Assumptions SEL_W and EX_W. For notational simplicity, we focus on Case (a) here; it is straightforward to draw analogous results for Case (b). With the existence of W, the MTR is defined as

$$\begin{aligned} m_d(u,w,x) &\equiv E[Y(d,w)|U=u,X=x] \\ &= \Pr[Y(d,w)=1|U=u,X=x], \end{aligned}$$

where W does not appear as a conditioning variable due to Assumption EX_W(a). Then, the target parameter can be expressed as

$$\tau = E[\tau_1(Z,W,X) - \tau_0(Z,W,X)],$$

where

$$\tau_d(z,w,x) = \int m_d(u,w,x) w_d(u,z,x) du.$$

Note that the weight $w_d(u, z, x)$ is not a function of w due to Assumption SEL_W(a).[5] Now, consider a mapping

$$(d, w) \mapsto y,$$

which is coded in the value e of $\epsilon \in \mathcal{E}$ where $|\mathcal{E}| = 16$ (redefining the variable ϵ introduced in Section 2.3). Conveniently, let $\mathcal{E} \equiv \{1, 2, ..., 16\}$; Table 2.1 lists all 16 maps. Equivalently, define

$$y = g_e(d, w),$$

which implies

$$Y(d, w) = g_\epsilon(d, w) \tag{2.5.1}$$

and $Y = g_\epsilon(D, W)$. By (2.5.1) and Assumption SEL_W(a), Assumption EX_W(a) can be equivalently stated as $(\epsilon, U) \perp (Z, W)|X$. Define the probability mass function of ϵ conditional on (U, X) as

$$q(e|u, x) \equiv \Pr[\epsilon = e|U = u, X = x] = \Pr[\epsilon = e|U = u, X = x, W = w].$$

Then, the MTR can be expressed as

$$m_d(u, w, x) = \sum_{e \in \mathcal{E}: g_e(d,w)=1} q(e|u, x) \tag{2.5.2}$$

[5] Apparently, with Assumption SEL_W(b), the weight will be written as $w_d^*(u, z, w, x)$ since the propensity score is a function of W.

ϵ	d	w	$Y(d,w)$	ϵ	d	w	$Y(d,w)$	ϵ	d	w	$Y(d,w)$	ϵ	d	w	$Y(d,w)$
1	0	0	0	5	0	0	0	9	0	0	0	13	0	0	0
	0	1	0		0	1	0		0	1	0		0	1	0
	1	0	0		1	0	1		1	0	0		1	0	1
	1	1	0		1	1	0		1	1	1		1	1	1
2	0	0	1	6	0	0	1	10	0	0	1	14	0	0	1
	0	1	0		0	1	0		0	1	0		0	1	0
	1	0	0		1	0	1		1	0	0		1	0	1
	1	1	0		1	1	0		1	1	1		1	1	1
3	0	0	0	7	0	0	0	11	0	0	0	15	0	0	0
	0	1	1		0	1	1		0	1	1		0	1	1
	1	0	0		1	0	1		1	0	0		1	0	1
	1	1	0		1	1	0		1	1	1		1	1	1
4	0	0	1	8	0	0	1	12	0	0	1	16	0	0	1
	0	1	1		0	1	1		0	1	1		0	1	1
	1	0	0		1	0	1		1	0	0		1	0	1
	1	1	0		1	1	0		1	1	1		1	1	1

Table 2.1All Possible Maps from (d,w) to y

and the target parameter as

$$\tau = \int E[\sum_{e:g_e(1,W)=1} q(e|u,X)w_1(u,Z,X)]du - \int E[\sum_{e:g_e(0,W)=1} q(e|u,X)w_0(u,Z,X)]du. \tag{2.5.3}$$

Recall, the sieve approximation of the conditional probability is given by $q(e|u,x) \approx \sum_{k\in\mathcal{K}} \theta_k^{e,x} b_k(u)$. Then, for $q \in \mathcal{Q}_K$, by (2.5.3), we have $\tau = E[\tau_1(Z,W,X)] - E[\tau_0(Z,W,X)]$ with

$$E[\tau_d(Z,W,X)] = \sum_{(w,x)\in\mathcal{W}\times\mathcal{X}} \sum_{e:g_e(d,w)=1} \sum_{k\in\mathcal{K}} \theta_k^{e,x} \gamma_k^d(w,x),$$

where $\gamma_k^d(w,x) \equiv \sum_{z\in\{0,1\}} p(z,w,x) \int b_k(u) w_d(u,z,x)du$ and $p(z,w,x) \equiv \Pr[Z = z, W = w, X = x]$. In terms of the data distribution, we can derive, e.g.,

$$
\begin{aligned}
p(y,1|z,w,x) &= \Pr[Y(1,w)=y, U \leq P(z,x)|X=x] \\
&= \int_0^{P(z,x)} \Pr[Y(1,w)=y|U=u,X=x]du \\
&= \sum_{e:g_e(1,w)=y} \int_0^{P(z,x)} q(e|u,x)du \\
&= \sum_{e:g_e(1,w)=y} \sum_{k\in\mathcal{K}} \theta_k^{e,x} \delta_k^1(z,x),
\end{aligned}
$$

where $\delta_k^d(z,x) \equiv \int_{\mathcal{U}_{z,x}^d} b_k(u)du$, as in the baseline case.

This modification yields a modified LP:

$$
\overline{\tau} = \max_{\theta\in\Theta_K} \sum_{(k,w,x)\in\mathcal{K}\times\mathcal{W}\times\mathcal{X}} \Big\{ \sum_{e:g_e(1,w)=1} \theta_k^{e,x}\gamma_k^1(w,x) - \sum_{e:g_e(0,w)=1} \theta_k^{e,x}\gamma_k^0(w,x) \Big\} \tag{LP$_W$1}
$$

$$
\underline{\tau} = \min_{\theta\in\Theta_K} \sum_{(k,w,x)\in\mathcal{K}\times\mathcal{W}\times\mathcal{X}} \Big\{ \sum_{e:g_e(1,w)=1} \theta_k^{e,x}\gamma_k^1(w,x) - \sum_{e:g_e(0,w)=1} \theta_k^{e,x}\gamma_k^0(w,x) \Big\} \tag{LP$_W$2}
$$

subject to

$$
\sum_{e:g_e(d,w)=1} \sum_{k\in\mathcal{K}} \theta_k^{e,x}\delta_k^d(z,x) = p(1,d|z,w,x) \qquad \forall (d,z,w,x) \in \{0,1\}^2 \times \mathcal{W}\times\mathcal{X}. \tag{LP$_W$3}
$$

Although omitted in the paper, similar modification can be made for Case (b), i.e., under Assumptions SEL$_W$(b) and EX$_W$(b). Since the number of maps increases to 16 instead of four of the baseline case, the dimension of the decision variable θ in the LP (LP$_W$1)–(LP$_W$3) is four times larger than that in the baseline. For example, assuming binary W and setting $K = 50$, we have

$dim(\theta) = 816 = 4 \times 204$. Still, it takes only about 13 seconds to solve the program.

We argue that in either Cases (a) or (b), the variation from W is in fact helpful in narrowing the bounds $[\underline{\tau}, \overline{\tau}]$ as long as W is a relevant variable. For the remainder of Section 2.6, we assume $W \in \{0, 1\}$ and suppress "conditional on X" for simplicity, unless otherwise noted.

Assumption R. *(i)* $\Pr[Y(d, w) \neq Y(d, w')] > 0$ *for some* d *and* $w \neq w'$*; (ii) either (a)* $P(z) > 0$ *for all* z *or (b)* $P(z, w) > 0$ *for all* z, w*.*

Theorem 2.5.1. *Under Assumptions* SEL_W*,* EX_W*, and R, the variation of* W *poses non-redundant constraints on* $\theta \in \Theta_K$ *in the LP* (LP_W1)–(LP_W3) *(suppressing* x*).*

Assumption R(i) is a relevance condition for W in determining Y. Heuristically, the improvement occurs because, with R(i), the constraint matrix (i.e., the matrix multiplied to the vector θ in (LP_W3)) has greater rank with the variation of W than without. See the proof of the theorem for a formal argument. Note that non-redundant constraints on θ do not always guarantee an improvement of the bounds in (LP_W1)–(LP_W3), because these constraints may still be non-binding. Nevertheless, non-redundancy is a necessary condition for the improvement.

2.6 Possible Identifying Assumptions

Bounds on the target parameter are generally uninformative in the absence of additional assumptions besides Assumptions SEL and EX. This is because a binary instrument has no extrapolative power for general non-compliers, e.g., always-takers and never-takers, but only identifies the effect for compliers. Prior studies have tried to overcome this challenge by imposing shape restrictions on the MTE (Thomas Cornelissen, Christian Dustmann, Anna Raute and Uta Schönberg (2016); Brinch, Mogstad and Wiswall (2017*a*); Kowalski (2020)), although these restrictions are not always empirically justified. Evidently, it would be useful to provide researchers with a larger variety of assumptions so that it is easier to find justifiable assumptions that suit their specific examples.

In Section 2.5, the existence of W (Assumptions SEL_W and EX_W) is shown to be one useful source for extrapolation. In this section, we propose identifying assumptions that can be incorporated within our framework and that help shrink the bounds on the target parameters. The shape restrictions employed in the literature can be used within our framework. We also propose other assumptions that have not been previously used in the LATE extrapolation. These assumptions can be incorporated as additional equality and inequality restrictions in the linear programming: Given the LP (∞-LP1)–

(∞-LP3), identifying assumptions can be imposed by appending

$$R_1 q = a_1, \qquad (\infty\text{-LP4})$$

$$R_2 q \leq a_2, \qquad (\infty\text{-LP5})$$

where R_1 and R_2 are linear operators on $\mathcal{Q}$ that correspond to equality and inequality constraints, respectively, and a_1 and a_2 are some vectors.

When an assumption violates the true data-generating process, then the identified set will be empty. This corresponds to the situation where the LP does not have a feasible solution. When we reflect sampling errors, this corresponds to the case where the confidence set is empty.[6]

2.6.1 Uniformity

Researchers may be willing to restrict the degree of treatment heterogeneity to yield informative bounds. This restriction has not been used before in the context of the MTE framework. This restriction may be combined with the assumptions related to the existence of W (Assumptions SEL_W, EX_W, and R). We suppress the conditioning on X throughout this subsection.

Assumption U. *For every* $w \in \mathcal{W}$, $\Pr[Y(1,w) \geq Y(0,w)] = 1$ *or* $\Pr[Y(1,w) \leq Y(0,w)] = 1$.

[6]In order to verify whether the identified set is empty, we need to check whether the feasible set of θ is empty. An efficient way to do this is to identify vertices of the feasible polytope, if any. This process is no simpler than the simplex algorithm that we use to solve the LP. Therefore, we recommend that one first solves the LP and check if infeasibility is reported.

When W is not available at all, this assumption can be understood with $\mathcal{W}$ being degenerate. The following assumption is stronger than Assumption U.

Assumption U*. *For every* $w, w' \in \mathcal{W}$, $\Pr[Y(1,w) \geq Y(0,w')] = 1$ *or* $\Pr[Y(1,w) \leq Y(0,w')] = 1$.

Note that w and w' may be the same or different, i.e., the uniformity is for all combinations of $(w, w') \in \{(0,0),(1,1),(1,0),(0,1)\}$. Therefore, Assumption U* implies Assumption U. Assumptions U and U* are weaker than the monotone treatment response assumption in Charles F Manski (1997) and Manski and Pepper (2000a) in that they do not impose the direction of monotonicity. Assumptions U and U* are also closely related to the rank similarity and rank invariance assumptions in the literature (e.g., Chernozhukov and Hansen (2005)). Namely, given a structural model $Y = 1[s(D,W) \geq V_D]$, when Assumption U* is violated, then rank similarity ($F_{V_1|U} = F_{V_0|U}$) cannot hold, and thus rank invariance ($V_1 = V_0$) cannot hold.

Assumptions U and U* can be imposed by "deactivating" relevant maps. For example, suppose $Y(1,w) \geq Y(0,w)$ almost surely for all $w \in \{0,1\}$ under Assumption U. This assumption can be imposed as equality constraints (∞-LP4), i.e., in the form of $R_1 q = a_1$, using the labeling of Table 2.1:

$$q(3|u) = q(4|u) = q(7|u) = q(8|u) = 0,$$
$$q(2|u) = q(4|u) = q(10|u) = q(12|u) = 0,$$

respectively, corresponding for $w = 1$ and $w = 0$. Therefore, the corresponding $\theta_k^e = 0$. Then, the effective dimension of θ will be reduced in (LP_W1)–(LP_W3) and thus yields narrower bounds. As another example, suppose the following holds almost surely under Assumption U*: $Y(1,1) \geq Y(0,0)$, $Y(1,0) \leq Y(0,1)$, $Y(1,1) \geq Y(0,1)$, and $Y(1,0) \geq Y(0,0)$. These inequalities respectively imply

$$
\begin{aligned}
&q(2|u) = q(4|u) = q(6|u) = q(8|u) = 0,\\
&q(5|u) = q(6|u) = q(13|u) = q(14|u) = 0,\\
&q(3|u) = q(4|u) = q(7|u) = q(8|u) = 0,\\
&q(2|u) = q(4|u) = q(10|u) = q(12|u) = 0.
\end{aligned}
$$

It is worth mentioning that, in Assumption U (Assumption U*), the direction of monotonicity is allowed to be different for different w ((w, w') pairs). This direction will be identified from the data. Specifically, the direction can be automatically determined from the LP by inspecting whether the LP has a feasible solution; when wrong maps are removed, there is no feasible solution. Note that this result holds regardless of the existence of W. It is easy to see that the direction of the monotonicity coincides with the sign of the ATE. Previous work has discussed the role of the rank similarity assumption on determining the sign of the ATE (Jay Bhattacharya, Azeem M Shaikh and Edward Vytlacil (2008*a*); Shaikh and Vytlacil (2011); Han (2020*b*)), and the result above shows that Assumptions U and U* play a similar role in the linear programming approach. In the next two subsections, we suppress W for

simplicity.

2.6.2 Direction of Endogeneity

In some applications, researchers are relatively confident about the direction of treatment endogeneity. The idea of imposing the direction of the selection bias as an identifying assumption appears in Manski and Pepper ($2000a$), who introduce monotone treatment selection (MTS), in addition to the monotone treatment response assumption mentioned above.

Assumption MTS. $E[Y(d)|D = 1, X = x] \geq E[Y(d)|D = 0, X = x]$ *for* $d \in \{0, 1\}$ *and* $x \in \mathcal{X}$.

Under our framework, this assumption can be imposed in the form of $R_2 q \leq a_2$. To see this, Assumption MTS is equivalent to

$$\sum_{e:g_e(d)=1} E\left[\int_{P(Z,X)}^{1} q(e|u)du - \int_{0}^{P(Z,X)} q(e|u)du \middle| X = x\right] \leq 0$$

for all $d, x \in \{0, 1\} \times \mathcal{X}$. As is clear from this expression, Assumption MTS imposes restrictions on the joint distribution of (ϵ, U).

2.6.3 Shape Restrictions

It is straightforward to incorporate the shape restrictions on the MTR or MTE function introduced in the literature. They can be imposed via constraints on θ.

Assumption M. *For* $x \in \mathcal{X}$, $m_d(u, x)$ *is weakly increasing in* $u \in [0, 1]$.

Assumption C. *For $x \in \mathcal{X}$, $m_d(u,x)$ is weakly concave in $u \in [0,1]$.*

Assumption M appears in Brinch, Mogstad and Wiswall (2017*a*) and Mogstad, Santos and Torgovitsky (2018) and Assumption C appears in Mogstad, Santos and Torgovitsky (2018). These assumptions can be imposed as inequality constraints (∞-LP4), i.e., in the form of $R_2 q \leq a_2$. For implications on the finite-dimensional LP (LP1)–(LP3), recall that for $q \in \mathcal{Q}_K$, the MTR satisfies

$$m_d(u,x) = \sum_{e:g_e(d)=1} q(e|u,x) = \sum_{k\in\mathcal{K}} \sum_{e:g_e(d)=1} \theta_k^{e,x} b_k(u).$$

According to the property of the Bernstein polynomial, Assumption M implies that $\sum_{e:g_e(d)=1} \theta_k^{e,x}$ is weakly increasing in k, i.e.,

$$\sum_{e:g_e(d)=1} \theta_1^{e,x} \leq \sum_{e:g_e(d)=1} \theta_2^{e,x} \leq \cdots \leq \sum_{e:g_e(d)=1} \theta_K^{e,x}.$$

Assumption C implies that

$$\sum_{e:g_e(d)=1} \theta_k^{e,x} - \sum_{e:g_e(d)=1} 2\theta_{k+1}^{e,x} + \sum_{e:g_e(d)=1} \theta_{k+2}^{e,x} \leq 0 \qquad \text{for } k = 0, ..., K-2.$$

One can obtain analogous assumptions and their implications in the presence of W.

Another shape restriction introduced in the literature is separability. Although it is not particularly appealing with binary Y, if one is willing to assume a separable model for $m_d(u,x) = \Pr[Y(d) = 1|U = u, X = x] = m_{1d}(x) + m_{2d}(u)$, then such a structure can be imposed on θ.

2.7 Simulation

This section provides numerical results to illustrate our theoretical framework and to show the role of different identifying assumptions in improving bounds on the target parameters. For target parameters, we consider the ATE and the LATEs for always-takers (LATE-AT), never-takers (LATE-NT), and compliers (LATE-C). We calculate the bounds on them based only on the information from the data and then show how additional assumptions on the existence of additional exogenous variables, uniformity, and shape restrictions tighten the bounds.

2.7.1 Data-Generating Process

We generate the observables (Y, D, Z, X, W) from the following data-generating process (DGP). We assume that W is a reverse IV, i.e., we maintain Assumptions SEL_W(a) and EX_W(a). We allow covariate X to be endogenous. All the variables are set to be binary with $\Pr[Z = 1] = 0.5$, $\Pr[X = 1] = 0.6$ and $\Pr[W = 1] = 0.4$. The treatment D is determined by Z and X through the threshold crossing model specified in Assumption SEL_W(a), where the propensity scores $P(z, x)$ are specified as follows: $P(0, 0) = 0.1$, $P(1, 0) = 0.4$, $P(0, 1) = 0.4$, and $P(1, 1) = 0.7$. The outcome Y is generated from (D, X, W) through $Y = DY_1 + (1 - D)Y_0$ where

$$Y_d = 1\left[m_d(U, X, W) \geq \epsilon\right] \tag{2.7.1}$$

and the MTR functions are defined as

$$m_0(u,0,0) = 0.02b_0^4(u) + 0.08b_1^4(u) + 0.14b_2^4(u) + 0.20b_3^4(u) + 0.21b_4^4(u)$$

$$m_1(u,0,0) = 0.12b_0^4(u) + 0.28b_1^4(u) + 0.44b_2^4(u) + 0.52b_3^4(u) + 0.54b_4^4(u)$$

$$m_0(u,1,0) = 0.22b_0^4(u) + 0.48b_1^4(u) + 0.64b_2^4(u) + 0.72b_3^4(u) + 0.74b_4^4(u)$$

$$m_1(u,1,0) = 0.44b_0^4(u) + 0.71b_1^4(u) + 0.88b_2^4(u) + 0.97b_3^4(u) + 0.99b_4^4(u)$$

$$m_0(u,0,1) = 0.10b_0^4(u) + 0.25b_1^4(u) + 0.30b_2^4(u) + 0.32b_3^4(u) + 0.33b_4^4(u)$$

$$m_1(u,0,1) = 0.20b_0^4(u) + 0.45b_1^4(u) + 0.60b_2^4(u) + 0.65b_3^4(u) + 0.66b_4^4(u)$$

$$m_0(u,1,1) = 0.30b_0^4(u) + 0.60b_1^4(u) + 0.80b_2^4(u) + 0.85b_3^4(u) + 0.86b_4^4(u)$$

$$m_1(u,1,1) = 0.35b_0^4(u) + 0.70b_1^4(u) + 0.92b_2^4(u) + 0.99b_3^4(u) + 1.00b_4^4(u)$$

where b_k^K stands for the k-th basis function in the Bernstein approximation of degree K. These MTR functions are chosen to be consistent with Assumptions M and C, i.e., to be positively monotone and weakly concave in u for all $(d,x,w) \in \{0,1\}^3$. Also, the DGP in (2.7.1) satisfies Assumption U* because ϵ does not depend on $d = 0,1$ and the MTR functions satisfy $m_1(u,x,w) > m_0(u,x,w)$ for all $(d,x,w) \in \{0,1\}^3$. Following the second example in Section 2.6.1, the DGP satisfy the following uniform order for the counterfactual outcomes $Y(d,w)$: $Y(1,1) \geq Y(0,1) \geq Y(1,0) \geq Y(0,0)$ a.s. We generate a sample containing 1,000,000 observations and choose $K = 50$. We choose the large sample size to mimic the population. Our choice of K is discussed below. The number of unknown parameters θ in the linear programming is equal to $\dim(\theta) = |\mathcal{E}| \times |\mathcal{X}| \times (K+1)$.

2.7.2 Bounds on Target Parameters under Different Assumptions

2.7.2.1 ATE

Figure 2.1 contains the bounds on the ATE under different assumptions. The true ATE value is 0.21, depicted as the solid red line in the figure. First, the worst-case bounds on the ATE with no additional assumptions (and without using variation from W) are $[-0.25, 0.45]$. Since the mappings do not involve W, we have $|\mathcal{E}| = 4$, and the linear programming is solved with $\dim(\theta) = |\mathcal{E}| \times |\mathcal{X}| \times (K+1) = 4 \times 2 \times 51 = 408$.

For comparison, we calculate the bounds that incorporate the existence of W. We build up the target parameters with mappings involving W and use data distribution conditional on $W = 0$ and $W = 1$ as the constraints. Using constraints conditional on different values of W allows us to fully exploit the variations from W; see ($\text{LP}_W 3$). As shown in the setup with W (Section 2.5), we have $|\mathcal{E}| = 16$, which gives $\dim(\theta) = |\mathcal{E}| \times |\mathcal{X}| \times (K+1) = 16 \times 2 \times 51 = 1,632$. The resulting bounds are depicted in the dotted greenish-blue line. When the variation from W is exploited, the bounds on the ATE are $[-0.24, 0.44]$, which is slightly narrower than without using W. This result is consistent with our theoretical finding presented in Theorem 2.5.1 that W can help tighten the bounds as long as it is a relevant variable. Nonetheless, these worst-case bounds are not that informative, e.g., they do not determine the sign of the ATE.

Next, we impose the uniformity assumption without W (Assumption U) and with W (Assumption U*). First, under Assumption U, the bounds on the

ATE are tightened as some mappings occur with probability zero reducing the dimension of θ. As mentioned in Section 2.6.1, the direction of monotonicity in Assumption U (i.e., which mapping does not occur) is determined by the LPs. We solve the LPs with different directions imposed, then choose the one with a feasible solution. This means that the corresponding direction of monotonicity is consistent with the DGP. Under Assumption U, we obtain a narrower bound $[0.06, 0.45]$. Second, under Assumption U*, the bounds become $[0.06, 0.33]$. In Figure 2.1, these bounds under Assumptions U and U* are depicted as violet and green dashed lines, respectively. Both sets of bounds identify the sign of the ATE, consistent with the theoretical discussion. While their lower bounds coincide, Assumption U* yields a lower upper bound compared to Assumption U.

Next, we impose the shape restrictions (Assumptions M and C). As discussed in Section 2.6.3, these assumptions can be easily incorporated in the linear programming by directly imposing inequality constraints on θ. Under these assumptions (and the existence of W), the bounds on the ATE shrink to $[0.11, 0.27]$, which is displayed with the pink line in Figure 2.1. We find that shape restrictions are powerful assumptions and yield narrower bounds compared to those with Assumption U*. They function differently in the linear programming: unlike the uniformity assumption, which maintains the ranking of individuals across counterfactual groups, shape restrictions directly control the MTR functions. Finally, the dash-dotted black line in Figure 2.1 shows the bounds on the ATE under the uniformity assumption and the

shape restrictions. These assumptions all together yield the narrowest bounds, $[0.13, 0.25]$, for the true ATE, 0.21.

2.7.2.2 Generalized LATEs

Next, we construct bounds on the generalized LATEs. The original definition of the LATE is the ATE for compliers (C). Researchers may also have interests in other local treatment effects. We consider two other parameters—LATEs for always-takers (AT) and never-takers (NT). Figure 2.2 displays the bounds on the LATE-AT, LATE-C, and LATE-NT under different assumptions. This analysis is analogous to that with the ATE. Since the covariate X affects the decision of compliance, to avoid confusion in the definition of the compliance groups, we instead establish bounds on the LATEs conditional on X. We draw the conditional MTE functions with solid red lines in both panels as a reference.

The feature that there exists no defiers in the DGP is known. When there is no defier, the LATE-C is point identified, which has an analytical expression of the two-stage least squares estimand. As a confirmation exercise, we numerically calculate the LATE-C using the linear programming, which yields point estimates as shown in Figure 2.2. The true LATE-Cs conditional on $X = 0$ and $X = 1$ are equal to 0.21 and 0.22, respectively. Regardless of assumptions imposed, the estimates remain close to the true values throughout.

The true values of the conditional LATE-AT and the LATE-NT are

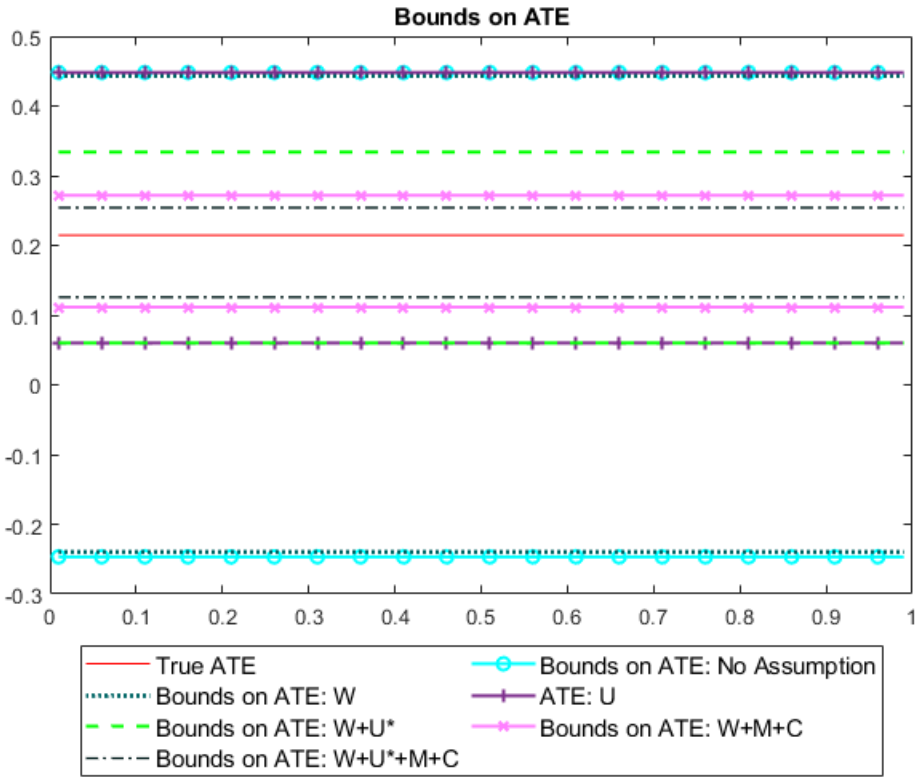


Figure 2.1. Bounds on the ATE under Different Assumptions

0.15 and 0.28 when $X = 0$ and 0.14 and 0.25 when $X = 1$. First, as before, we consider the worst-case bounds where the existence of W is ignored versus where W is taken into account. Without W, we get the bounds $[-0.71, 0.24]$ and $[-0.28, 0.72]$ on the LATE-AT and the LATE-NT conditional on $X = 0$, and $[-0.48, 0.52]$ and $[-0.56, 0.43]$ conditional on $X = 1$; with W, we get the bounds $[-0.62, 0.2]$ and $[-0.28, 0.7]$ on the LATE-AT and the LATE-NT conditional on $X = 0$, and $[-0.48, 0.5]$ and $[-0.55, 0.41]$ conditional on $X = 1$. The upper bounds with W are lower than the ones without W, although the gain is not substantial. For the lower bounds, the one on the LATE-AT conditional on $X = 0$ is significantly higher with W than without W, and all the other ones have negligible differences with and without W.

We then apply Assumptions U and U*. Under Assumption U, the bounds on the LATE-AT and the LATE-NT turn to $[0, 0.24]$ and $[0, 0.72]$ conditional on $X = 0$, and $[0, 0.52]$ and $[0, 0.43]$ conditional on $X = 1$; when W is used and Assumption U* is applied, the bounds shrink to $[0, 0.18]$ and $[0, 0.47]$ conditional on $X = 0$, and $[0, 0.36]$ and $[0, 0.35]$ conditional on $X = 1$. As before, Assumptions U and U* determine the sign of the effects.

When the shape restrictions are imposed instead, the bounds on the LATE-AT and the LATE-NT were improved to $[0.11, 0.17]$ and $[0.03, 0.3]$ conditional on $X = 0$, and $[0.05, 0.31]$ and $[0.15, 0.31]$ conditional on $X = 1$. Under Assumption U* combined with the shape restrictions, we get the narrowest bounds of $[0.11, 0.15]$ and $[0.04, 0.3]$ conditional on $X = 0$, and $[0.08, 0.26]$ and $[0.15, 0.31]$ conditional on $X = 1$.

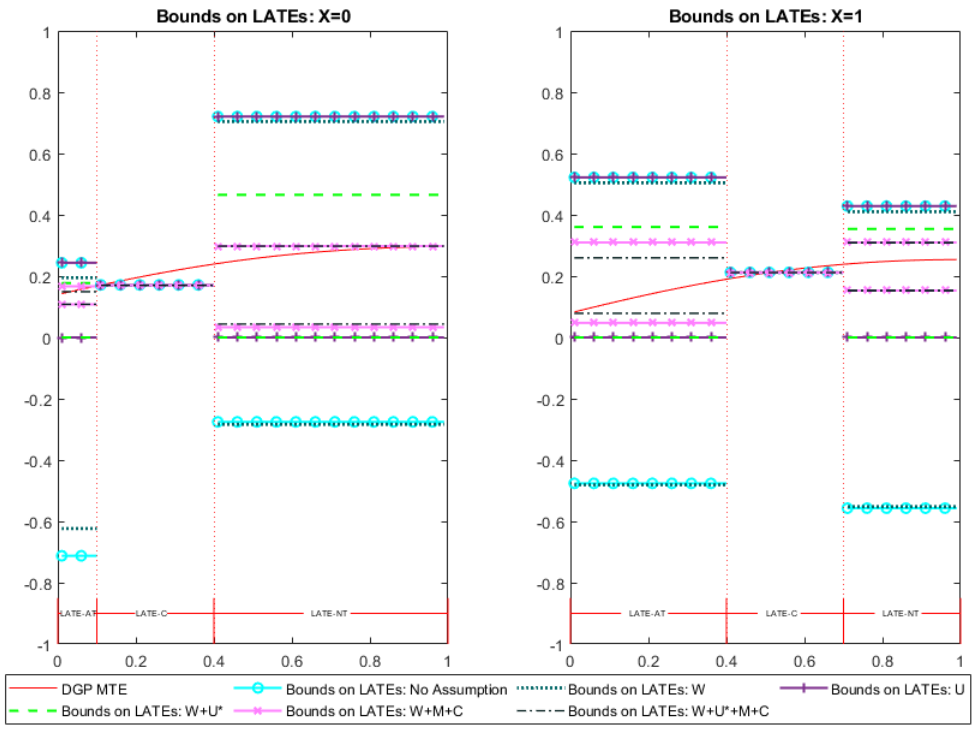


Figure 2.2. Bounds on the LATEs under Different Assumptions

2.7.3 The Choice of K

As a tuning parameter in the LP, we need to choose the order of Bernstein polynomials, K. In general, K should be chosen based on the sample size and the smoothness of the function to be approximated, in our case, $q(\cdot)$. The choice of the sieve dimension or more generally, regularization parameters, is a difficult question (Xiaohong Chen (2007)) and developing data-driven procedure is a subject of on-going research in various nonparametric contexts of point identification; see, e.g., Xiaohong Chen and Timothy Christensen (2015) and Sukjin Han (2020*a*). In this partial identification setup, we propose the following heuristic and conservative approach, which is in spirit consistent with the very motivation of partial identification.

First, we do not want to claim any prior knowledge about the smoothness of $q(\cdot)$ because it is the distribution of a latent variable. Because K determines the dimension of unknown parameter θ in the linear programming, the width of the bounds tends to increase with K. At the same time, the computational burden increases with K. One interesting numerical finding is that, when K is sufficiently large, the increase of the width slows down and the bounds become stable. This suggests that we may be able to conservatively choose K that acknowledges our lack of knowledge of the smoothness but, at the same time, produces a reasonable computational task for the linear programming.

To illustrate this point, we consider the conditional MTE and ATE as the target parameters and show how their bounds change as we increase K. We

consider the MTE because it is a fundamental parameter that generates other target parameters, and hence, it is important to understand the sensitivity of its bounds to K. Figures 2.3 and 2.4 show the evolution of the bounds on the MTE and the ATE as K grows. When $K = 5$, the bounds are narrow. Although it may be tempting to choose this value of K, this attempt should be avoided as it may be subject to the misspecification of smoothness. When K increases beyond 30, the bounds start to converge and become stable. We choose $K = 50$, and this is the choice we made in our previous numerical exercises.[7]

As discussed in Section B.2.1 in the Appendix, it is worth mentioning that the bounds on the MTE are point-wise sharp but *not* uniformly sharp. The graph for the MTE bounds are drawn by calculating the point-wise sharp bounds on MTE at each point of u (after properly discretizing it) and then connecting them. Therefore, these bounds should *not* be viewed as uniformly sharp bounds. Nonetheless, this graph is still useful for the purpose of our illustration. Given the current DGP, we find that there are no uniformly sharp bounds for the MTE.

[7]Note that with larger K, some LP solvers would ignore coefficients with negligible (e.g., 10^{-13}) values that cause a large range of magnitude in the coefficient matrix. It may be recommended to simultaneously rescale a column and a row to achieve a smaller range in the coefficients.

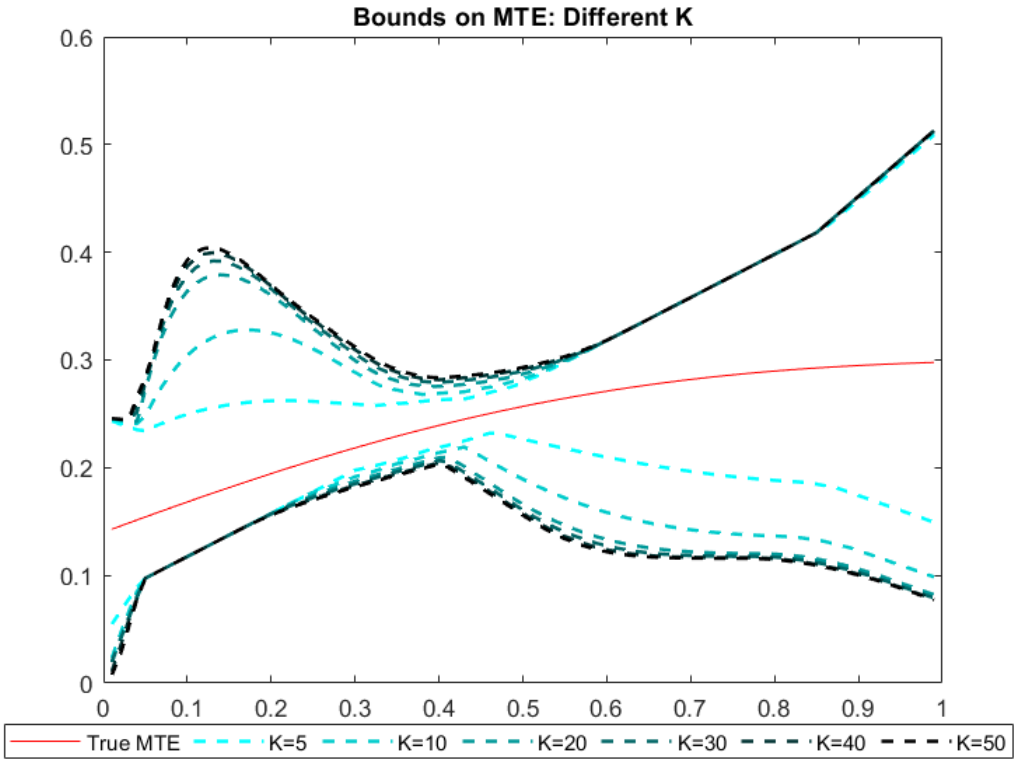


Figure 2.3. Bounds on MTE with Different K

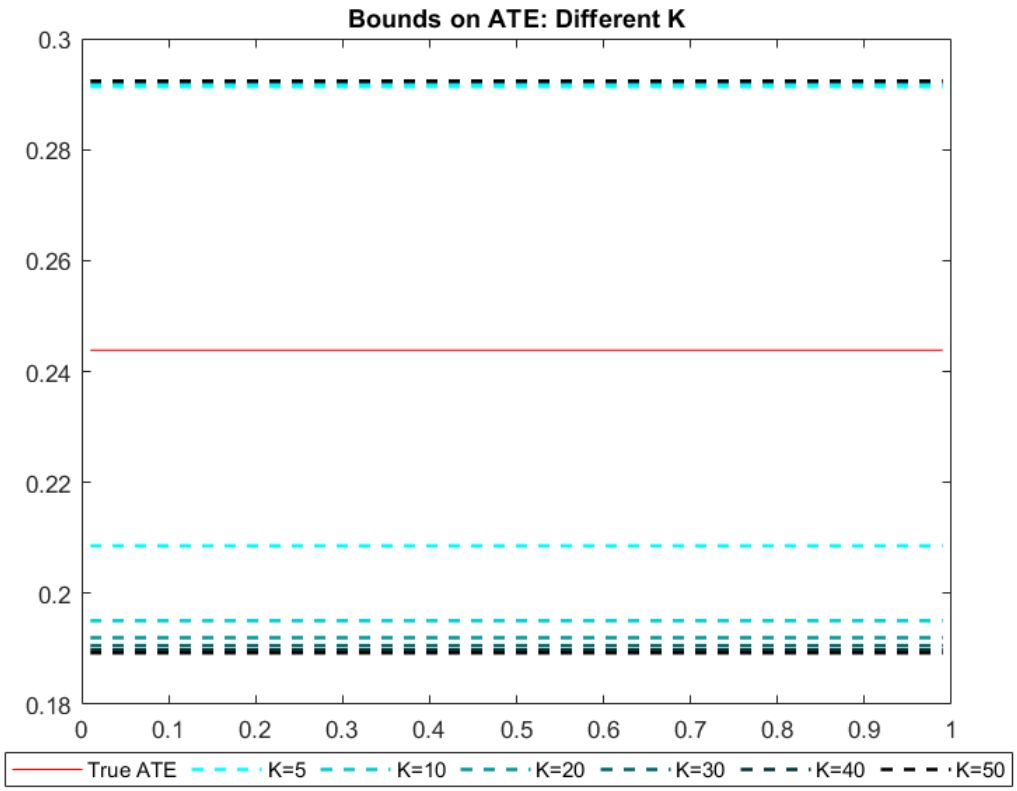


Figure 2.4. Bounds on ATE with Different K

2.8 Empirical Application

It is widely recognized in the empirical literature that health insurance coverage can be an essential factor for the utilization of medical services (Michael D Hurd and Kathleen McGarry (1997); Dorothy D Dunlop, Larry M Manheim, Jing Song and Rowland W Chang (2002); Finkelstein et al. (2012); Sarah L Taubman, Heidi L Allen, Bill J Wright, Katherine Baicker and Amy N Finkelstein (2014)). Prior studies on this topic typically make use of parametric econometric models for the analysis. In their application, Han and Lee (2019) relax this common approach by introducing a semiparametric bivariate probit model to measure the average effect of insurance coverage on patients' medical visits. By applying our theoretical framework of partial identification, we further relax the parametric and semiparametric structures used in these studies. More importantly, we try to understand how much we can learn about the effect of insurance that is utilized through various counterfactual policies by learning the effect of different compliance groups.

We use the 2010 wave of the Medical Expenditure Panel Survey (MEPS) and focus on all the medical visits in January 2010. The sample is restricted to contain individuals aged between 25 and 64 and exclude those who had any kind of federal or state insurance in 2010. The outcome Y is a binary variable indicating whether or not an individual has visited a doctor's office; the treatment D is whether an individual has private insurance. We choose whether a firm has multiple locations as the binary instrument Z. This IV reflects the size of the firm, and larger firms are more likely to provide fringe benefits, in-

cluding health insurance. On the other hand, the number of branches of a firm does not directly affect employee decisions about medical visits. To justify the IV, self-employed individuals are excluded. For potentially endogenous covariates X, we include the age being 45 and older, gender, income above median, and health condition. Lastly, for an exogenous covariate W, we use the percentage of workers who are provided with paid sick leave benefits within each industry. Following Han and Lee (2019), we assume W satisfies Assumptions $\text{SEL}_W(\text{b})$ and $\text{EX}_W(\text{b})$, as X is controlled. We construct a categorical variable such that $W = 0$ for less than 50%, $W = 1$ for between 50–80%, and $W = 2$ for above 80%. Table 2.2 summarizes the observables.

Table 2.2Summary Statistics

	Variable	Mean	S.D	Min	Max
Y	Whether or not visit doctors	0.18	0.39	0	1
D	Whether or not have insurance	0.66	0.47	0	1
Z	Firm has multiple locations	0.68	0.47	0	1
X	Age above 45	0.41	0.49	0	1
	Gender	0.50	0.50	0	1
	Income above median	0.50	0.50	0	1
	Good health	0.36	0.48	0	1
W	Pay sick leave provision	1.25	0.73	0	2
Number of observations = 7,555					

First, as a benchmark, we report that the LATE-C estimate calculated via our linear programming approach is equal to a singleton of 0.17, which is in fact identical to the 2SLS estimate we separately calculate. In what follows, we extrapolate this LATE beyond the complier group to the ATE. The

presence of covariates reduces the effective sample size and thus leads to larger sampling errors in estimating the p of the ∞-LP (∞-LP1)–(∞-LP3). This may create inconsistencies in the set of equality constraints (∞-LP3), resulting in no feasible solution. This is in fact what happens in this application. To resolve this estimation problem, we introduce a slackness parameter η and modify (∞-LP3) so that, with some slackness, it satisfies

$$\|R_0 q - p\| \leq \eta. \tag{2.8.1}$$

A similarly modified constraint can then be followed in the finite-dimensional LP after approximation, as well as by combining (∞-LP4)–(∞-LP5). The appropriate value of η should depend on the sample size, the dimension of covariates, and the dimension of the unknown parameter θ. To explain the latter, as K increases, the dimension of θ (i.e., unknowns) increases, while the number of constraints (i.e., simultaneous equations for the unknowns) is fixed. Therefore, as K increases, the chance that the LP does not have a feasible solution would decrease. Based on the method discussed in the previous section, we set $K = 50$ in this application.

We calculate worst-case bounds on the ATE, as well as bounds after imposing Assumptions U and M and after using covariate W. Under Assumption U, the data rules out the possibility that $Y(0) > Y(1)$, indicating that individuals with private insurance are more likely to visit a doctor. Assumption M imposes that the MTR function is weakly increasing in $U = u$. Usually, U is interpreted as the latent cost of obtaining treatment. Kowalski (2020)

interpreted U as eligibility in a similar setup for Medicaid insurance. The eligibility for Medicaid is related to income level and age. In our setup, because the treatment is having the private insurance, we interpret the eligibility as the health status, which is reflected in the premium. Interpreting U as a latent cost (e.g., premium) of getting private insurance, Assumption M states that the chance of making a medical visit (with or without insurance) increases for those with higher cost. This is a reasonable assumption given that sicker individuals typically face higher insurance costs and also visit doctors more often. We choose the slackness parameter η to be 0.05 under no assumption and Assumption U and 0.07 when Assumption M is added. When W is used, we choose η to be 0.08 under no assumption and 0.1 with Assumption M.

The bounds on the ATE are shown in Figure 2.5. The worst-case bound on the ATE equals $[-0.45, 0.37]$. The bounds become $[0.01, 0.37]$ under Assumption U and $[0.06, 0.37]$ under Assumption M. It is interesting to note that the identifying power of the uniformity and the shape restriction is similar in this example. When both Assumption U and Assumption M are imposed, the bounds are further tightened to $[0.07, 0.37]$, although not substantially, indicating that the two assumptions are complementary. Lastly, we see improvements when the variation in W is exploited than when it is not, although the gains are not large.

Next, we consider the always-taker, complier, and never-taker LATEs. We consider these generalized LATEs conditional on $X = x$. Specifically, we focus on the treatment effects for males above age 45, with income below the

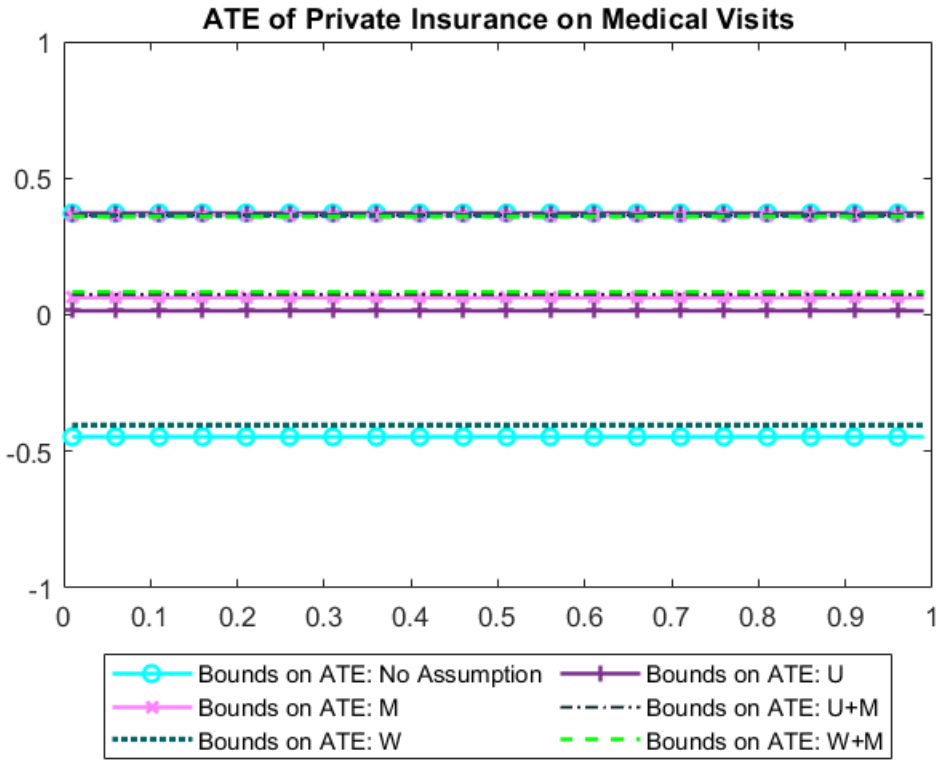


Figure 2.5. Bounds on the ATE of Private Insurance on Medical Visits

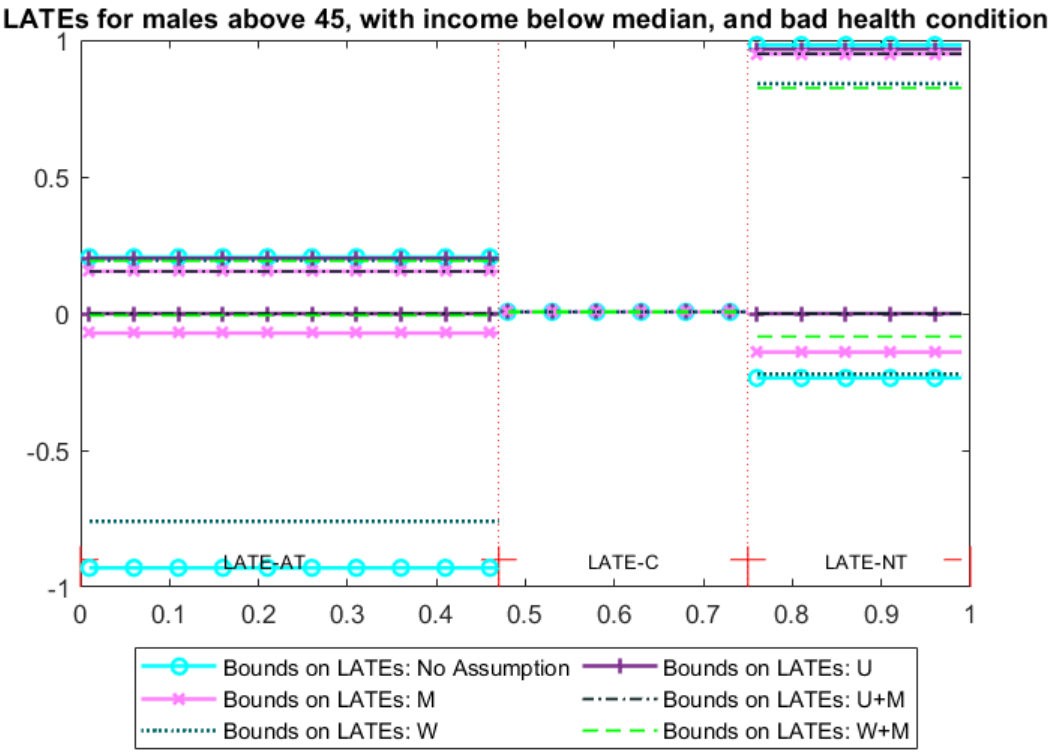


Figure 2.6. Bounds on the generalized LATEs of Private Insurance on Medical Visits for Male Above 45, with Income Below Median, of Bad Healthiness

median and bad health conditions. The results are shown in Table 2.3 and depicted in Figure 2.6. The LATE-C is analytically calculated via TSLS.[8] For the LATE-AT and LATE-NT, Assumption U identifies the sign of the effects, and Assumption M nearly identifies it. Using the variation in W mostly improves the bounds compared to the ones without it. From the results, we can conclude that possessing private insurance has the greatest effect on medical visits for never takers, i.e., people who face higher insurance cost. This provides a policy implication that lowering the cost of private insurance is important, because high costs might hinder those with the most need from receiving enough medical services.

[8]When the alternative constraint (2.8.1) is used with the slackness parameter, the LATE-C is no longer a singleton.

Table 2.3Estimated Bounds on generalized LATEs for Males Above 45, with Income Below Median, Bad Health Condition

	No Assumption	U	M	U+M	W	M+W
LATE-AT	[-0.93,0.21]	[0,0.20]	[-0.07,0.15]	[0,0.15]	[-0.76,0.20]	[-0.01,0.19]
LATE-C	0.01	0.01	0.01	0.01	0.01	0.01
LATE-NT	[-0.24,0.98]	[0,0.97]	[-0.14,0.95]	[0,0.95]	[-0.22,0.84]	[-0.08,0.82]
Slackness parameter η	0.05	0.05	0.07	0.07	0.08	0.10
Number of observations = 7,555						

Chapter 3

Partial Identification on Treatment Effect on Transitions and Its Empirical Application

This paper focuses on identifying the treatment effect on conditional transition rate in a discrete-time range when the outcome of interest is a transition from an initial state to a destination state. Treatment is assigned only once at time zero or at a later time. Even when the treatment is randomly assigned, it is challenging to identify the treatment effect due to selective dropouts over time, referred to as dynamic selection. Previous literature dealing with this problem usually imposes parametric or semi-parametric assumptions. This paper tries to solve the dynamic selection problem in treatment effects estimation on conditional transition probabilities under a non-parametric setting. It gives a tighter partial identification bound with randomization assumption, comparing to the bound obtained in Johan Vikström, Geert Ridder and Martin Weidner (2018) and argues it is the tightest possible bound under randomization assumption. Then this paper relaxes the random assumption under the condition mean independence assumption.

This paper mainly focuses on the situation where treatment is assigned during the initial period. Analogous results can be reached in a case where the treatment is assigned at a later time. When the assignment is random, the

treatment effect is point identified in the initial period. However, in the following periods, as people leave the initial state, the fraction remaining in the treatment and control groups starts to differ, and the survivors have different characteristics. Thus, the randomization assumption does not help in point identification after the first period. Vikström, Ridder and Weidner (2018)proposes partial identification bounds on treatment effect under the initial random assignment assumption by taking advantage of information from survivors in the control group to identify the counterfactual. Their bound is not sharp because the endpoints can only be reached under two additional restrictions. Moreover, when an absorbing state is allowed, information from non-survivors in the control group can be combined to give bounds on the counterfactual. Mathematically, define Y_t as the realized outcome and $\overline{Y}_{t-1}$ as a vector of realized historical outcomes, the bound given in this paper takes effect when $P(Y_t = 1|\overline{Y}_{t-1} \neq \vec{0}, D = 0)P(\overline{Y}_{t-1} \neq \vec{0}|D = 0) > P(Y_t = 1|\overline{Y}_{t-1} = \vec{0}, D = 0)P(\overline{Y}_{t-1} = \vec{0}|D = 0)$. The intuition behind this inequality is that when the fraction of non-survivors is large so that they give more various current outcomes, it helps in shrinking the original bound.

When the randomization assumption is relaxed, challenges in estimation lie in both endogenous selection into treatment and selective dropout. This chapter shows that under the conditional mean independence assumption, bounds on the treatment effect after the first period can be captured with endogenously chosen treatment. This bound can further be tightened under the Monotone Treatment Selection Assumption and the Monotone Treatment

Response Selection. With Monotone Treatment Selection Assumption, this result can also be extended to identify quantile differences when the outcome is a mixed random variable.

Vikström, Ridder and Weidner (2018) analyzes the effect of the Illinois unemployment bonus experiment using the partial identification bound. However, I find almost zero difference when comparing bound in this paper to the original bound, up to tiny differences in the second period. To show the improvement brought by the new bound this paper proposes under randomization assumption, we conduct some numerical exercises with a duration model. It shows how the new bound takes functions with the variation of different parameters with graphs showing both the original bound and the new bound. With the non-randomization assumption, I use the partial identification bound to estimate the labor market return of a college degree with NLSY79 data. Binary labor market returns chosen here are employment status above the median income. The results show a wide range of effects of a college degree on employment status but narrower bounds of an increasing effect on income being above the median under the conditional mean independence assumption. Pre-believes of initial selection direction or treatment effect direction narrow the bound. Incomes are used as a mixed outcome variable in quantile difference analysis. From the results, when outcome distribution under treatment first order dominates outcome distribution under non-treatment, the differences in quantiles are mostly non-positive. When we assume the reverse, the difference in quantiles is positive and significant. These results give

some views contradictory to commonly believed significant education return. However, the minor variation in data limits functions of partial identification bounds. These bounds can be used as a range of possible treatment effects to test results under stronger assumptions.

The remainder of the paper consists of five parts. Section I gives motivation, background, and literature review. Section II defines the treatment effect and constructs identification under randomization. Section III gives identification without random assignment. Section IV conducts numerical exercises under randomization and shows results from applying bounds obtained in Section III. Section V concludes and discusses future work that needs to be done.

3.1 Motivation and Literature Review

Consider an intervention when the outcome is a transition from an initial state to a destination state. In labor policy research, if the initial state is unemployment and the destination state is re-employment, economists are interested in the effect of an intervention such as an unemployment benefits program or a training program. Because this transition does not happen instantaneously, the evolution of the treatment effect is a vital interest, which will be referred to as the dynamic treatment effect here. Another parameter of interest in a similar situation is the treatment effect on the CDF of duration time to transit. Using this definition of treatment effect avoids dynamic selection problems and captures the effect of the treatment on the average duration

or survival time. However, using this definition does not reveal the evolution path of treatment effects. Taking a medical experiment as an example, two new drugs have the same effect on the average survival time, but one of them has a constant effect over time, and another one has a minimal effect at the beginning but a more substantial effect in following periods, pharmacists may value them differently. In economics study, the close relationship between hazard rates and micro theory indicates that gaining knowledge of the evolution path of treatment effect helps test economic predictions.

Challenge in estimating the dynamic treatment effect on the conditional transition rate lies in the dynamic selection problem. Dynamic selection exists even with randomly assigned treatment. This selection comes from the possibility that different fractions of people in the treatment and control group leave the initial state in each period. Another possibility is that people leaving the initial state in the treatment group are from a different portion of characteristics distribution, comparing to those who leave from the control group, which causes endogenous selection in each period even when fraction leaving the initial period are the same in two groups. Suppose there is a job training program aiming to help the unemployed get re-hired. In the first period, unemployed individuals are randomly assigned to a job training program. Workers in the treated group are more likely to be hired in the first period if this program positively helps to gain specific job skills. Then in the second period, those remaining unemployed in the treatment group might be from a lower tail in ability distribution comparing with the control group. This possibility indi-

cates potential biases if we identify the dynamic treatment effect by merely comparing the realized outcomes of the treatment group and control group. Previous literature dealing with this problem usually imposes parametric or semi-parametric assumptions and focuses on identifying parameters in a structural model with an endogenous treatment. One of the most frequently used models is the proportional hazard (PH) model and the mixed proportional hazard (MPH) model. The proportional hazard model imposes a multiplicative assumption, where they write the instantaneous transition rate as a product of time effect, intervention effect, and an unobservable individual effect. Chris Elbers and Geert Ridder (1982) showed the MPH model could be nonparametrically identified. In the empirical literature, Bruce D Meyer (1996) used the PH model to estimate the effect evolution of the Illinois Unemployment Bonus Experiment and showed longer experiment spell increases the unemployment rate, indicating the undesirability of a permanent program. Jaap H Abbring and Gerard J Van den Berg (2003) used the MPH model to estimate the duration model. They used a continuous-time range and showed non-identification under the non-parametric assumption. The other way to resolve the dynamic selection problem is the threshold crossing model, constructed by James J Heckman and Salvador Navarro (2007), where they forced the existence of covariates uncorrelated with unobservables and non-recurrent states, and one of the major focuses is on dynamic discrete choice model identification. Other literature on similar topic includes Flavio Cunha, James Heckman and Salvador Navarro (2005), Thierry Magnac and David Thesmar (2002) etc.

Vikström, Ridder and Weidner (2018) gives partial identification when semi-parametric models do not hold. They first give partial identification bounds on average treatment effect on survivors in each period, under solely randomization assumption. To tighten that bound, they further impose three additional economic assumptions and use them on the Illinois Unemployment Bonus Experiment. They find the re-employment bonus effect was increasing over time and had a sharp increase before the bonus claim deadline. After the deadline, there was no effect. Moreover, this effect was heterogeneous across racial and income ranking groups. This finding is consistent with Meyer (1996) and labor supply prediction. Besides, their results can coexist with recurrent states and heterogeneous treatment effects.

This paper is based on Vikström, Ridder and Weidner (2018). Under the same setup and the randomization assumption, I tighten the bounds using the extra information obtained from data and argue the combination of these two bounds gives a sharp bound. Then using a similar method, I relax the randomization assumption under conditional mean independence assumption. This paper will supplement the literature in dealing with dynamic selection problems in dynamic treatment estimation in a discrete-time range and can be used to test the performance of parametric/semi-parametric point estimation results.

3.2 Partial Identification on Average Treatment Effect on Treated Survivors Under Randomization Assumption

3.2.1 Set Up

This paper focuses on treatment effects on transition rates in discrete time. Treatment is assigned at time 0 and transitions occur at times $t = 1, 2, ...$ Let potential outcome Y_t^1 be the indicator of a transition in period t if treated, and Y_t^0 be the potential outcome if not treated. The observed outcome is:

$$Y_t = DY_t^1 + (1 - D)Y_t^0 \tag{3.2.1}$$

Here absorbing states are allowed. Therefore, after getting to the destination state, returning to the initial state in later periods is still possible. This follows Vikström, Ridder and Weidner (2018). Absorbing state is a special case under generalized identification. The parameter of interest in this paper is a dynamic analog of the average treatment effect on the treated in the static setting:

Definition 1. *The causal effect on the transition probability of the treated survivors in t is the Average Treatment Effect on Treated Survivors (ATETS) defined by:*

$$ATETS_t = E(Y_t^1 | Y_{t-1}^1 = 0, ..., Y_1^1 = 0) - E(Y_t^0 | Y_{t-1}^1 = 0, ..., Y_1^1 = 0)$$

This definition follows Vikström, Ridder and Weidner (2018). To correctly define the dynamic treatment effect, one needs to condition the same dynamic history, which makes it more difficult to define the average treatment

effect than the static case. The ATETS captures how the presence of a treatment benefits/harms an individual, and for simplicity, it will be referred to as treatment effect in the following sections.

3.2.2 Identification Under Random Assignment

Assumption 1 (Random assignment of treatment)**.**

$$D \perp\!\!\!\perp \{Y_t^1, Y_t^0 : t = 1, 2, ...\}$$

In Vickström et al. (2016), they constructed partial identification bounds of equation (1) under Assumption 1. The result is listed as Proposition 1.

Proposition 1. [1] *Suppose Assumption 1 holds:*

For $t = 1$:

$$ATETS_t = E(Y_1^1) - E(Y_1^0) = E(Y_1|D = 1) - E(Y_1|D = 0) \tag{3.2.2}$$

For $t \in \{2, 3, 4, ...\}$, define $\overline{Y}_{t-1} = (Y_{t-1}, Y_{t-2}, ..., Y_1)$ which represents a vector of realized outcomes in each history period. If $P(\overline{Y}_{t-1} = 0|D = 1) = 0$, then $ATETS_t$ is not defined; if $P(\overline{Y}_{t-1} = 0|D = 1) > 0$, and also $P(D = 1) > 0$,

[1] This Proposition comes from Vikström, Ridder and Weidner (2018) Theorem 1. Proof can be found in Vikström, Ridder and Weidner (2018) Appendix.

$P(D=0)>0$, then we have the bounds:$LB_t \leq ATETS_t \leq UB_t$, where

$$\begin{aligned} LB_t \equiv & P(Y_t=1|\overline{Y}_{t-1}=0,D=1) \\ & - min\{1, \frac{1-[1-P(Y_t=1|\overline{Y}_{t-1}=0,D=0)]P(\overline{Y}_{t-1}=0|D=0)}{P(\overline{Y}_{t-1}=0|D=1)}\} \end{aligned} \tag{3.2.3}$$

$$\begin{aligned} UB_t \equiv & P(Y_t=1|\overline{Y}_{t-1}=0,D=1) \\ & - max\{0, 1+\frac{P(Y_t=1|\overline{Y}_{t-1}=0,D=0)P(\overline{Y}_{t-1}|D=0)-1}{P(\overline{Y}_{t-1}=0|D=1)}\} \end{aligned} \tag{3.2.4}$$

Using the definition from Elie Tamer (2010), a bound is sharp when any value in the set, including the end points, cannot be rejected as the true value under current assumptions. The assumptions given in Vikström, Ridder and Weidner (2018) are random assignment of treatment and $P(\overline{Y}_{t-1}=0|D=1)>0$. However, when they calculate the bound on $ATETS_t$, they start by giving bounds on $p^0(1|0,\neq 0)$ and $p^0(1|\neq 0,0)$. This means they obtain the bound in Theorem 1 under extra restrictions that:

$$p_{t-1}(0,\neq 0) = Pr(\overline{Y}^1_{t-1}=0, \overline{Y}^0_{t-1}\neq 0) \neq 0$$

$$p_{t-1}(\neq 0,0) = Pr(\overline{Y}^1_{t-1}\neq 0, \overline{Y}^0_{t-1}= 0) \neq 0$$

To clarify the differences between the bound in Vikström, Ridder and Weidner (2018) and the later bound, I call this bound the first bound. Under these two restrictions, the first bound is the tightest possible bound. There are two ways to show it. First, as stated in Corollary 1 in Vikström, Ridder and Weidner (2018), the lower bound and upper bound equal to each other if both $P(\overline{Y}_{t-1}=0|D=0)=1$ and $P(\overline{Y}_{t-1}=0|D=1)=1$, which can be

realized under existing assumptions. Second, the ends of the first bound can be achieved when $p_{t-1}(0,0) = max\{Pr(\overline{Y}^1_{t-1} = 0) + Pr(\overline{Y}^0_{t-1} = 0) - 1, 0\}$; $p^0(1|0, \neq 0) = 1$, $p^0(1| \neq 0, 0) = 0$ or $p^0(1|0, \neq 0) = 0$, $p^0(1| \neq 0, 0) = 1$ These equalities can hold together. However, when we allow $p_{t-1}(0, \neq 0) = 0$ and $p_{t-1}(\neq 0, 0)$, the first bound is dysfunctional, and the end points can be rejected as a true value. Therefore, by considering cases allowing $p_{t-1}(0, \neq 0) = 0$ or $p_{t-1}(\neq 0, 0)$, the first bound can be tightened.

Following same procedure used in Vikström, Ridder and Weidner (2018), if we start by giving bounds on $p^0(1|0, 0)$ and $p^0(1| \neq 0, \neq 0)$, we can get another bound allowing $p_{t-1}(0, \neq 0) = 0$ or ${}_{t-1}(\neq 0, 0) = 0$, but requiring $p_{t-1}(0, 0) \neq 0$ and $p_{t-1}(\neq 0, \neq 0) \neq 0$. We can call it the second bound. These two bounds are supplementary to each other. When none of $p_{t-1}(0, 0)$, $p_{t-1}(0, \neq 0)$, $p_{t-1}(\neq 0, 0)$, $p_{t-1}(\neq 0, \neq 0)$ equals to 0, both bounds have effect on giving boundaries to $ATETS_t$. If one of them equals to 0, say $p_{t-1}(0, 0) = 0$ or $p_{t-1}(\neq 0, \neq 0) = 0$, it can be shown that the second bound is large enough so that only the first bound is effective, since restrictions used in the second bound is violated. Symmetrically, if $p_{t-1}(0, \neq 0) = 0$ or $p_{t-1}(\neq 0, 0) = 0$, the first bound gives worst case bound and the second bound takes effect. Thus taking the intersection of them gives a tighter bound, and because this bound covers all potential cases, the boundary can be reached requiring no extra assumption, we can argue the intersection of the first bound and the second bound is sharp.

Theorem 1. *Suppose Assumption 1 holds, for $t = 1$:*

$$ATETS_t = E(Y_1^1) - E(Y_1^0) = E(Y_1|D = 1) - E(Y_1|D = 0) \tag{3.2.5}$$

For $t = 2, 3, 4, ...$, if $P(\overline{Y}_{t-1} = 0|D = 1) = 0$, then $ATETS_t$ is not defined; if $P(\overline{Y}_{t-1} = 0|D = 1) > 0$, and also $P(D = 1) > 0$, $P(D = 0) > 0$, then we have the bounds: $LB'_t \leq ATETS_t \leq UB'_t$, where

$$LB'_t \equiv P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 1) - \frac{min\{P(Y_t = 1|D = 0), P(\overline{Y}_{t-1} = 0|D = 1)\}}{P(\overline{Y}_{t-1} = 0|D = 1)} \tag{3.2.6}$$

$$UB'_t \equiv P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 1) - \frac{max\{0, P(Y_t = 1|D = 0) + P(\overline{Y}_{t-1} = 0|D = 1) - 1\}}{P(\overline{Y}_{t-1} = 0|D = 1)} \tag{3.2.7}$$

Lemma 1. *The bound proposed in Theorem 1 is sharp 1.*

Proof of Theorem 1 and Lemma 1 follow directly from Hoeffding inequality and Fréchet-Hoeffding theorem. As in Proposition 1, results in Theorem 1 does not require assumptions beyond random assignment and $P(\overline{Y}_{t-1} = 0|D = 1) > 0$. However, one important condition for Theorem 1 to give tighter bound is an allowance for recurrent state. If absorbing state is assumed instead, results in Theorem 1 will give the same bounds as in Proposition 1. Assuming recurrent state, bound from Theorem 1 is tighter when

$$\begin{aligned} &P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0)P(\overline{Y}_{t-1} = 0|D = 0) \\ &\quad < P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)P(\overline{Y}_{t-1} \neq 0|D = 0) \end{aligned}$$

To give an interpretation of this condition, first consider intuition behind the bound obtained in Theorem 1. First, the bound is increasing in $P(Y_t =$

$1|\overline{Y}_{t-1} = 0, D = 1)$; this follows that the parameter of interest here is average treatment effect of treated survivors; thus a higher realized transition rate indicates potentially larger treatment effect. Then, since higher transition rate among control group indicates potentially smaller effect of treatment, both the upper bound and lower bound are decreasing in $P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0)$ and $P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)$. Formally, we can write the counterfactual part as:

$$P(Y_t^0 = 1|\overline{Y}_{t-1}^1 = 0) = \frac{p_t^0(1|0,0)p_{t-1}(0,0) + p_t^0(1|0,\neq 0)p_{t-1}(0,\neq 0)}{p_{t-1}(0,0) + p_{t-1}(0,\neq 0)}$$

Using law of total probability and make substitutions, we can write it as:

$$\begin{aligned}
&P(Y_t^0 = 1|\overline{Y}_{t-1}^1 = 0) \\
&= \frac{P(\overline{Y}_{t-1} = 0|D = 0)P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0) - p_t^0(1|\neq 0, 0)p_{t-1}(\neq 0, 0)}{P(\overline{Y}_{t-1} = 0|D = 1)} \\
&+ \frac{P(\overline{Y}_{t-1} \neq 0|D = 0)P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0) - p_t^0(1|\neq 0, \neq 0)p_{t-1}(\neq 0, \neq 0)}{P(\overline{Y}_{t-1} = 0|D = 1)}
\end{aligned}$$

From above, the counterfactual is increasing in $P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0)$ and $P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)$. The interpretation behind it is that transition rate on survivors and non-survivors in the control group both give information on their potential outcome in the treatment group. Vikström, Ridder and Weidner (2018) uses the transition rate of survivors in the control group and imposing bound on the fraction of individuals who would be survivors under both treatment and non-treatment, i.e., $p_{t-1}(0,0)$ and gives a bound on their potential outcome under treatment. Results in Theorem 1 tighten it by using both information from survivors and non-survivors in the control group.

Another important determinant of the bound is the fraction of survivors. If $P(\overline{Y}_{t-1}|D=1)$ and $P(\overline{Y}_{t-1}|D=0)$ are small, difference between lower bound and upper bound is larger, since smaller observation pool drives less variation.

Corollary 1 shows that if $P(\overline{Y}_{t-1}=0|D=1)=1$ and $P(\overline{Y}_{t-1}=0|D=0)=1$ or 0, $ATETS_t$ is point identified.

Corollary 1. *$ATETS_t$ is point identified if $P(\overline{Y}_{t-1}=0|D=1)=1$, and $P(\overline{Y}_{t-1}=0|D=0)=1$ or 0.*

If $P(\overline{Y}_{t-1}=0|D=1)=1$ and $P(\overline{Y}_{t-1}=0|D=0)=1$,

$$ATETS_t = P(Y_t=1|\overline{Y}_{t-1}=0, D=1) \tag{3.2.8}$$

If $P(\overline{Y}_{t-1}=0|D=1)=1$ and $P(\overline{Y}_{t-1}=0|D=0)=0$,

$$ATETS_t = P(Y_t=1|\overline{Y}_{t-1}=0, D=1) - P(Y_t=1|\overline{Y}_{t-1}\neq 0, D=0) \tag{3.2.9}$$

3.2.3 Bounds under Additional Assumptions

This section corresponds to Section 4 in Vikström, Ridder and Weidner (2018), where they discuss improvement on bound obtained in Proposition 1 under application of three additional assumptions-Monotone Treatment Response (MTR), Common Shock (CS) and Positively Correlated Outcomes (PCO). Recall under randomization assumption, the authors derive bound by imposing bounds on $p_t^0(1|0,\neq 0)$ and $p_t^0(1|\neq 0,0)$ in step one and then imposing bounds on $p_{t-1}(0,0)$ in step two. MTR assumption combining with CS assumption gives a tighter bound on $p_{t-1}(0,0)$, therefore tightens the original

bound. However, it does not helps in tightening bound on $p_{t-1}(0, \neq 0)$, so it is unable to shrink the second bound using these assumptions. The possibly tightest bound under these assumptions would be an intersection of the first bound under MTR+CS assumption and the second bound. In contrast, PCO assumption tightens the bound by giving more precise set in the first step. It narrows sets on $p_t^1(1|0, \neq 0)$ and $p_t^1(1| \neq 0, 0)$, which is not used in defining the second bound. However, using the same spirit, we can give a stronger assumption such that positively correlated outcomes exist under conditional on any potential history instead of only $\overline{Y}_{t-1}^1 = 0$ and $\overline{Y}_{t-1}^0 = 0$. This stronger assumption gives a small improvement. A more efficient way will be discussed in the future to incorporate the PCO assumption better.

3.3 Partial Identification on Average Treatment Effect on Treated Survivors Under Non-Random Assignment

Results from the previous section can be informative in analyzing experimental results. However, in economics research, randomization is a strong assumption. To generalize the results from Theorem 1, a similar method in Vikström, Ridder and Weidner (2018) is used in this section such that the randomization assumption can be relaxed.

3.3.1 Identification under Non-Random Assignment

Without random treatment assignment assumption, Assumption 1 does not hold, thus another assumption capturing the relationship between potential outcomes and realized outcomes is required. The key assumption in this section is a conditional mean independence assumption. This assumption assumes conditional on realized history outcomes; the current outcome is mean independent of counterfactual history outcomes.

Assumption 2 (Conditional Mean Independence)**.** *Define* Y_t^1 *as outcome in period* t *with treatment and* Y_t^0 *as outcome in period* t *without treatment.* $\overline{Y}_{t-1}^1 = (Y_1^1, Y_2^1, ..., Y_{t-1}^1)$, $\overline{Y}_{t-1}^0 = (Y_1^0, Y_2^0, ..., Y_{t-1}^0)$ *are vector of historical outcomes with treatment and without treatment respectively. We have,* $\forall t$, $\forall$ $\overline{Y}_{t-1}^1$ *and* $\overline{Y}_{t-1}^0$,

$$
\begin{aligned}
E(Y_t^1|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=1) &= E(Y_t^1|\overline{Y}_{t-1}^1, D=1) \\
E(Y_t^0|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=0) &= E(Y_t^0|\overline{Y}_{t-1}^0, D=0)
\end{aligned}
$$

This assumption assumes the realized past outcomes completely capture intertemporal dependence. It can be realized when the treatment group and control group receive different random shocks. To give a better interpretation of this assumption, a discrete duration model is used as an example:

$$
\begin{aligned}
Y_{it}^0 &= I(\alpha_t + V_i - \epsilon_{it}^0 \geq 0) \\
Y_{it}^1 &= I(\alpha_t + \gamma_{it} + V_i - \epsilon_{it}^1 \geq 0) \\
D &= I\Big[\sum_{t=1}^{\infty} \frac{1}{(1+\rho_i)^t}(E(U_{it}(Y_{it}^1)) - E(U_{it}(Y_{it}^0)) \geq 0\Big]
\end{aligned}
\tag{3.3.1}
$$

In the model above, α_t is a fixed time effect; V_i is unobserved heterogeneity; ϵ_{it}^0 and ϵ_{it}^1 are random shocks received by control group and treatment group respectively, and they are assumed to be independent. Intertemporal dependence of random shocks are allowed in this model.

Without randomization assumption, a decision rule needs to be specified. I assume people choose to get treatment if the discounted summation of expected utility gain from treatment is non-negative. Here ρ_i is a discounting factor, measuring weight put on each period by individual i. U_{it} is time-individual specific utility function. Because the outcome is binary, when we assume U_{it} is monotone for any i and t, we can replace the decision rule by:

$$D_i = I\Big[\sum_{t=1}^{\infty} \frac{1}{(1+\rho_i)^t}(EY_{it}^1 - EY_{it}^0) \geq 0\Big]$$

Under assumption that $\{\epsilon_{it}^1\} \perp\!\!\!\perp \{\epsilon_{it}^0\}$,

$$\begin{aligned}
&E(Y_t^1|\overline{Y}_{t-1}^1 = \vec{a}, \overline{Y}_{t-1}^0 = \vec{b}; V_i, D=1) \\
&= Pr(\epsilon_{it}^1 \leq \alpha_t + \gamma_{it} + V_i | \epsilon_{it-1}^1 \in A_{it-1}, ..., \epsilon_{i1}^1 \in A_{i1}; \epsilon_{it-1}^0 \in B_{it-1}, ..., \epsilon_{i1}^0 \in B_{i1}; V_i, D=1) \\
&= Pr(\epsilon_{it}^1 \leq \alpha_t + \gamma_{it} + V_i | \epsilon_{it-1}^1 \in A_{t-1}, ..., \epsilon_{i1}^1 \in A_1; V_i, D=1) \\
&= E(Y_t^1|\overline{Y}_{t-1}^1 = \vec{a}, V_i, D=1)
\end{aligned}$$

where $\vec{a}$ and $\vec{b}$ are $t-1$ dimensional vector composing of 0 and 1. When $\epsilon_{ij}^1 \in A_{ij}, j = 1, ..., t-1, \overline{Y}_j^1 = \vec{a}_j$. Definition of B_{ij} is symmetric. $E(Y_t^0|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D = 0) = E(Y_t^0|\overline{Y}_{t-1}^0, D=0)$ can be showed in the same way. Because this hold for all V_i, we can drop it, thus, $E(Y_t^1|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=1) = E(Y_t^1|\overline{Y}_{t-1}^1, D=1)$. Under Assumption 2, $ATETS_t$ can be partially identified.

Theorem 2. *Suppose Assumption 2 holds. Let* $t \in \{2, 3, 4, ...\}$*, If* $P(\overline{Y}_{t-1}|D = 1)P(D = 1) + P(D = 0) = 0$*, $ATETS_t$ is not well defined. If* $P(\overline{Y}_{t-1}|D = 1)P(D = 1) + P(D = 0) > 0$*, then we have*

$$LB_t \leq ATETS_t \leq UB_t$$

where

$$LB_t = -\Bigg\{ \frac{P(D=1)P(\overline{Y}_{t-1}=0|D=1)[1 - P(Y_t=1|\overline{Y}_{t-1}=0, D=1)]}{P(\overline{Y}_{t-1}=0|D=1)P(D=1) + P(D=0)} + \frac{P(D=0)P(Y_t=1|D=0)}{P(\overline{Y}_{t-1}=0|D=1)P(D=1) + P(D=0)} \Bigg\} \tag{3.3.2}$$

$$UB_t = \frac{P(D=1)P(\overline{Y}_{t-1}=0|D=1)P(Y_t=1|\overline{Y}_{t-1}=0, D=1)}{P(\overline{Y}_{t-1}=0|D=1)P(D=1) + P(D=0)} + max\Bigg\{ \frac{P(D=0)P(Y_t=1|\overline{Y}_{t-1}=0, D=1)}{P(\overline{Y}_{t-1}=0|D=1)P(D=1) + P(D=0)}, \frac{P(D=0)(1 - min\{P(Y_t=1|\overline{Y}_{t-1}=0, D=0)), P(Y_t=1|\overline{Y}_{t-1}\neq 0, D=0)\}}{P(\overline{Y}_{t-1}=0|D=1)P(D=1) + P(D=0)} \Bigg\} \tag{3.3.3}$$

Proof of Theorem 2 is in Appendix.

Note this bound displays similar features discussed in Section 1, LB_t and UB_t are both increasing in $P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 1)$, and decreasing in $P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0)$ and $P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)$, and length of the bound is decreasing in fraction of survivors in treatment and control group since larger fraction makes variation more possible.

So far this paper considers situations where the treatment is assigned at time zero. The results in Theorem 1 and Theorem 2 can be extended

to cases where treatment is assigned at a later time. Following notations in Vikström, Ridder and Weidner (2018), denote indicator of a transition in period t if the treatment started in period $k \leq t$ as Y_t^k, and assume treatment is an absorbing state. Use $ATETS_t(k)$ to denote treatment effect at time t of treatment assigned at time $k \leq t$, and focus on survivors firstly treated at time t, we can derive bounds on $ATETS_t(k)$ applying results in Theorem 1 and Theorem 2, by treating k as the first period, and condition changes from $\overline{Y}_{t-1} = 0$ to $\overline{Y}_{t-1}^k = 0, \overline{D}_{k-1} = 0$, where $\overline{D}_{k-1} = (D_{k-1}, D_{K-2}, ...D_0)$.

3.3.2 Bounds under Additional Assumptions

This subsection gives some examples of how additional assumptions can be added to tight the bound proposed by Theorem 2. Without non-random assumption, one can make assumptions on the selection process, the direction of the outcome, and the correlation between potential outcomes to give a narrower bound. Assumption 2 can be seen as an assumption on the relationship between potential outcomes. Therefore, this section focuses on two classical assumptions restricting selection direction and outcome direction. The first assumption made is Monotone Treatment Selection (MTS) Assumption, which is proposed and used by Charles F Manski and John V Pepper (2000*b*) and assumes people choosing to be treated get better (or worse) potential outcome on average comparing with people choosing to not to be treated. It is equivalent to first order stochastic dominance assumption by Richard Blundell, Amanda Gosling, Hidehiko Ichimura and Costas Meghir (2007). MTR assumption is

also used in Vikström, Ridder and Weidner (2018) Section 4 and it assumes the positive or negative effect of treatment for all observations. Without a prior, MTS and MTR assumptions separately give the same bound as bound in Theorem 3. When giving a prior, MTS and MTR assumptions give tighter bounds.

Assumption 3 (Monotone Treatment Selection (MTS)). *For all* $d = 0, 1$, *and* $(\overline{Y}^1_{t-1}, \overline{Y}^0_{t-1}) \in \{0,1\}^{2(t-1)}$, *we have either:*

$$\textit{Positive MTS:} \quad E(Y^d_t | \overline{Y}^1_{t-1}, \overline{Y}^0_{t-1}, D = 1) \geq E(Y^d_t | \overline{Y}^1_{t-1}, \overline{Y}^0_{t-1}, D = 0)$$

or

$$\textit{Positive MTS:} \quad E(Y^d_t | \overline{Y}^1_{t-1}, \overline{Y}^0_{t-1}, D = 1) \leq E(Y^d_t | \overline{Y}^1_{t-1}, \overline{Y}^0_{t-1}, D = 0)$$

Assumption 4 (Monotone Treatment Response (MTR)). *Define* $H_t = (\overline{Y}^1_t, \overline{Y}^0_t)$ *as a vector of potential historical outcomes.Then Either*

$$\textit{Positive MTR:} \quad E(Y^1_t | H_{t-1}, D) \geq E(Y^0_t | H_{t-1}, D)$$

or

$$\textit{Negative MTR:} \quad E(Y^1_t | H_{t-1}, D) \leq E(Y^0_t | H_{t-1}, D)$$

$\forall t$, $\forall H_{t-1}$, $\forall D$.

Combining MTS and MTR assumption gives a stronger assumption that selection into treatment is of the same direction of treatment effect. Under this assumption, if the treatment has positive effect on all observations, the ones with outcome-friendly characteristics choose to get treated; If the

treatment has negative effect, they will avoid being treated. This assumption is consistent with both utility maximization problem with complete information and game with asymmetric information in micro theory. Under complete information, it can be assumed that the unobserved characteristics are positively correlated with both potential outcome and budget. Under incomplete information, this assumption means people with characteristics benefiting potential outcome have more information on the treatment so they can choose to take or avoid the treatment. Based on these stories, I use "Selection Power" to denote this assumption.

Assumption 5 (Selection Power (SP))**.** *Suppose Assumption 3 and Assumption 4 hold, we have:*

$$E(Y_t^1|H_{t-1},D) \geq E(Y_t^0|H_{t-1},D), \forall t, D \Rightarrow E(Y_t^d|H_{t-1},D=1) \geq E(Y_t^d|H_{t-1},D=0), \forall d,t$$

$$E(Y_t^1|H_{t-1},D) \leq E(Y_t^0|H_{t-1},D), \forall t, D \Rightarrow E(Y_t^d|H_{t-1},D=1) \leq E(Y_t^d|H_{t-1},D=0), \forall d,t$$

These assumptions can be incorporated into the duration model given by equation (10). MTR assumption can be simply obtained by imposing sign restriction on γ_{it}, with positive γ_{it} indicating positive treatment response and negative γ_{it} indicating the reverse. Assumption 5 is a special case of Assumption 3, I give an example to incorporate both Assumption 3 and Assumption 5. To do this, we need to be specific on distribution of ϵ_{it}^0 and ϵ_{it}^1. To simplify calculation, I assume $\gamma_{it} = \gamma_t f(V_i)$ such that heterogeneous treatment effect is

captured by a multiplication of common treatment effect r_t and a function of unobserved heterogeneity. Thus the potential outcome turns into:

$$Y_{it}^0 = I(\alpha_t + V_i - \epsilon_{it}^0 \geq 0)$$

$$Y_{it}^1 = I(\alpha_t + \gamma_t f(V_i) + V_i - \epsilon_{it}^1 \geq 0)$$

Then I make assumption on distribution of random shocks such that

$$\epsilon_{it}^1 \sim logistic(\gamma_{it}\overline{V}, 1), \quad \epsilon_{it}^0 \sim logistic(0, 1)$$

Here I assume that shock received by treatment group is related with heterogeneous treatment effect and average characteristics $\overline{V}$. If we assume V_i captures ability, then this assumption can be interpreted as that one's outcome under treatment is affected by level of her ability comparing with the average level. Thus one's decision on treatment is decided by her knowledge of her unobserved heterogeneity. Under these assumptions, the decision rule can be written as:

$$D = I\Big[\sum_{t=0}^{\infty} \frac{1}{(1+\rho)^t} \cdot \frac{e^{-\alpha_t - V_i}(1 - e^{\gamma_t(f(\overline{V}) - f(V_i))})}{(1 + e^{-\alpha_t - g(V_i) + \gamma(f(\overline{V}) - f(V_i))})(1 + e^{-\alpha_t - V_i})} \geq 0\Big]$$

Thus the decision on treatment is determined by the sign of $1 - e^{\gamma_t(f(\overline{V}) - f(V_i))}$. If we assume then function $f(\cdot)$ is increasing and $\gamma_t > 0$, $\forall t$, then $D = 1$ only when $V_i \geq \overline{V}$ and $D = 0$ when $V_i < \overline{V}$. Thus:

$$\begin{aligned} P(Y_t^d = 1 | \overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D = 1) &= P(Y_t^1 = 1 | \overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, V_i \geq \overline{V}) \\ &> P(Y_t^d = 1 | \overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, V_i < \overline{V}) = P(Y_t^d = 1 | \overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D = 0) \end{aligned}$$

The results under Assumption 3 to 5 are listed in Theorem 3 to 5.

Theorem 3 (Bounds on ATETS under MTS)**.** *Suppose that Assumption 2 and Assumption 3 hold. Let LB_t and UB_t be the lower bound and upper bound in Theorem 3, then we have:*

$$\begin{aligned} LB_t \leq ATETS_t \leq & P(Y_t|\overline{Y}_{t-1}=0, D=1) \\ & - min\{P(Y_t=1|\overline{Y}_{t-1}=0, D=0), P(Y_t=1|\overline{Y}_{t-1}\neq 0, D=0)\} \end{aligned} \tag{3.3.4}$$

under positive MTS, and

$$\begin{aligned} UB_t \geq ATETS_t \geq & P(Y_t|\overline{Y}_{t-1}=0, D=1) \\ & - max\{P(Y_t=1|\overline{Y}_{t-1}=0, D=0), P(Y_t=1|\overline{Y}_{t-1}\neq 0, D=0)\} \end{aligned} \tag{3.3.5}$$

under negative MTS.

Theorem 4 (Bounds on ATETS under MTR)**.** *Suppose that Assumption 2 and Assumption 4 hold. Let LB_t and UB_t be the lower bound and upper bound in Theorem 3, then we have:*

$$0 \leq ATETS_t \leq UB_t \text{ under positive MTR} \tag{3.3.6}$$

and

$$LB_t \geq ATETS_t \geq 0 \text{ under negative MTR.} \tag{3.3.7}$$

Theorem 5 (Bounds on ATETS under SP)**.** *Suppose that Assumption 2 and Assumption 5 hold, then we have:*

$$\begin{aligned} 0 \leq ATETS_t \leq & P(Y_t|\overline{Y}_{t-1}=0, D=1) \\ & - min\{P(Y_t=1|\overline{Y}_{t-1}=0, D=0), P(Y_t=1|\overline{Y}_{t-1}\neq 0, D=0)\} \end{aligned} \tag{3.3.8}$$

under positive MTS and MTR, and

$$0 \geq ATETS_t \geq P(Y_t|\overline{Y}_{t-1}=0, D=1) - max\{P(Y_t=1|\overline{Y}_{t-1}=0, D=0), P(Y_t=1|\overline{Y}_{t-1}\neq 0, D=0)\} \tag{3.3.9}$$

under negative MTS and MTR.

3.3.3 Mixed Outcome

The results above can be extended to cases where the outcome is a mixed variable. Since this paper researches on the performance of treatment on survivors in each period, the outcome is required to have a positive possibility at zero and it can be continuous at least in some feasible interval. $ATETS_t$ is not able to be identified in this scenario. Instead, we can identify differences in distribution, which can be captured by differences in percentiles. By identifying bounds on conditional CDF in each period, Charles F Manski and C Sims (1994) provides a way to give bounds on the difference on quantiles between treatment group and control group. To make identification on bounds on CDF, we need to make modification on Assumption 2 and make it a stronger assumption.

Assumption 6 (Conditional Independence)**.** *Define* $Y_t^1 \geq 0$ *as a mixed variable with positive probability at 0 in period t with treatment and* $Y_t^0 \geq 0$ *as outcome in period t without treatment.* $\overline{Y}_{t-1}^1 = (Y_1^1, Y_2^1, ..., Y_{t-1}^1)$, $\overline{Y}_{t-1}^0 = (Y_1^0, Y_2^0, ..., Y_{t-1}^0)$ *are vector of historical outcomes with treatment and without treatment respectively,* $F(\cdot|\cdot)$ *is conditional CDF. We have,* $\forall t$, $\forall$ $\overline{Y}_{t-1}^1$ *and* $\overline{Y}_{t-1}^0$,

$$F_{Y_t^1}(y_t^1|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=1) = F_{Y_t^1}(y_t^1|\overline{Y}_{t-1}^1, D=1)$$

$$F_{Y_t^0}(y_t^0|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=0) = F_{Y_t^0}(y_t^0|\overline{Y}_{t-1}^0, D=0)$$

Assumption 6 solely cannot be used to identify bounds on conditional CDF. Thus, this paper combines it with First Order Stochastic Dominance Assumption and shows under these two assumptions, one side of quantile difference can be identified.

Assumption 7 (First Order Stochastic Dominance)**.** $\forall\, d = 0, 1$ *and* $\forall(\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0)$,

$$F_{Y_t^d}(y_t^d|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=1) \leq F_{Y_t^d}(y_t^d|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=0)$$

or

$$E_{Y_t^d}(y_t^d|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=1) \geq F_{Y_t^d}(y_t^d|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=0)$$

Theorem 6. *Suppose Assumption 6 and Assumption 7 hold. For* $d = 0, 1$ *define the* $\alpha - quantile of P(Y_t^d|\overline{Y}_{t-1}^1 = 0)$ *as:*

$$q^d(\alpha|\overline{Y}_{t-1}^1 = 0) \equiv min\{\omega \in \Omega : P(Y_t^d \leq \omega|\overline{Y}_{t-1}^1 = 0) \geq \alpha\}$$

If $F_{Y_t^d}(y_t^d|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=1)$ *first order dominates* $F_{Y_t^d}(y_t^d|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=0)$,

$$q^1(\alpha|\overline{Y}_{t-1}^1 = 0) - q^0(\alpha|\overline{Y}_{t-1}^1 = 0) \leq s(\alpha) \tag{3.3.10}$$

where

$s(\alpha) \equiv inf\{\omega_1 \in \Omega : P(Y_t \leq \omega_1|\overline{Y}_{t-1} = 0, D=1) \geq \alpha\}$

$- inf\{\omega_0 \in \Omega : max\{P(Y_t \leq \omega_0|\overline{Y}_{t-1} = 0, D=0), P(Y_t \leq \omega_0|\overline{Y}_{t-1} \neq 0, D=0)\} > 1-\alpha\}$

If $F_{Y_t^d}(y_t^d|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=0)$ *first order dominates* $F_{Y_t^d}(y_t^d|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=1)$,

$$q^1(\alpha|\overline{Y}_{t-1}^1 = 0) - q^0(\alpha|\overline{Y}_{t-1}^1 = 0) \geq r(\alpha) \tag{3.3.11}$$

where

$$r(\alpha) \equiv inf\{\omega_0 \in \Omega : P(Y_t \leq \omega_1|\overline{Y}_{t-1} = 0, D=1) \geq 1-\alpha\}$$
$$- inf\{\omega_0 \in \Omega : min\{P(Y_t \leq \omega_0|\overline{Y}_{t-1} = 0, D=0), P(Y_t \leq \omega_0|\overline{Y}_{t-1} \neq 0, D=0)\} > \alpha\}$$

3.4 Application

3.4.1 Numerical Exercises under Random Assumption

To show improvement induced by Theorem 1, I conduct some numerical exercises using the duration model of the outcome given by equation (10), we have:

$$Y_{it}^0 = I(\alpha_t + V_i - \epsilon_{it}^0 \geq 0)$$

$$Y_{it}^1 = I(\alpha_t + \gamma_{it} + V_i - \epsilon_{it}^1 \geq 0)$$

Now assume a two period model following the duration model above, where V_i is a one-dimensional characteristic lying in$[-5, 5]$ and $\gamma_{it} = \gamma_t$. ϵ_{it}^d is assumed to follow AR(1) process and $\epsilon_{it}^1 \perp\!\!\!\perp \epsilon_{it}^0$. Assume:

$$\epsilon_{i1}^0 \sim N(0,1), \qquad \epsilon_{i1}^1 \sim logit(0,1)$$

$$\epsilon_{i2}^0 = \epsilon_{i1}^0 + N(0,1) \quad \epsilon_{i2}^1 = \epsilon_{i1}^1 + N(0,1)$$

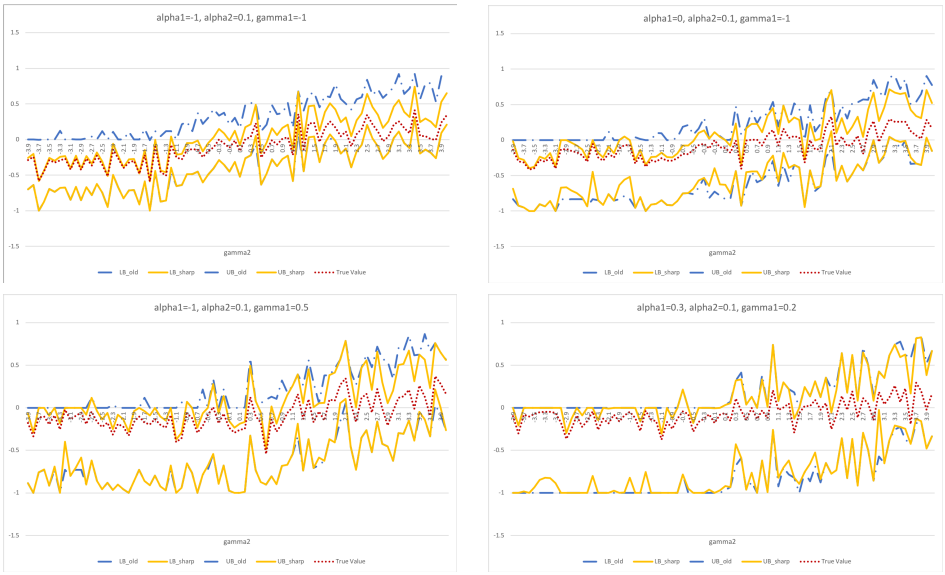
$ATETS_2$ can be seen as a reduced form approximation of γ_2, to show how results from Theorem 1 tightens bound in Proposition 1, I give graphs how the two bounds differ across true values of $\gamma_2 \in [-0.9, 0.9]$, given values of γ_1, α_1 and α_2 fixed. The results are shown in Figure 3.1. The dashed lines show results from Theorem 1 and solid lines show results from Proposition 1.

From Figure 3.1., all lower bounds and upper bounds are increasing with the true value of γ_2, showing partial identification bounds capture features of true treatment effect-driven parameters. However, as the true value of γ_2 increases, the bounds on $ATETS_t$ turn into less precise approximation. Figure 3.1 gives another fact that when $\gamma_1 < 0$, result from Theorem 1 has the most significant difference from the previous result. Recall that the bound in Proposition 1 is obtained by making restriction that $p_{t-1}(\neq 0, 0) \neq 0$. When γ_1 is negative, it means non-survivors in treatment group lie right to non-survivors in the control group on the distribution of V_i. Since survivors in have a lower level of V_i comparing with non-survivors, the difference between V_i of survivors in the control group and non-survivors in the treatment group. Therefore, situation where $p_{t-1}(\neq 0, 0) = 0$ is more possible. Thus, allowing $p_{t-1}(\neq 0, 0) = 0$ gives a tighter bound. Influences from fixed time effects on the difference between two bounds are not obvious, and this is because both treatment group and control group receive same fixed time effect; thus the effect of fixed time effect on selective dropout is small.

3.4.2 Under Non-Random Assumption: Labor Market Return of College Diploma

It is a common belief that workers with a diploma of advanced education have better economic outcomes comparing with others. This prediction is consistent with both the signaling theory and human capital accumulation theory. However, because of dynamic selection, it is not clear how does this effect change over time. Knowledge of treatment effect evolution path also

Figure 3.1. Comparison Between Results from Proposition 1 and Theorem 1



helps in distinguishing the mechanism behind the effect of a college degree. Assuming the treatment effect is entirely from human capital accumulation, then we would expect to see a stable treatment effect over time. If it is not the case, we would predict that a college diploma has at least some signaling value on the labor market.

Data used in this section is from NLSY79. Here The treatment group had attained a college degree in 1979, while the control group did not have a college degree in 1979 and did not get it until 1990. Two labor market outcomes are considered: employment status and income above median. In researches on educational return, positive MTR and positive MTS are usually assumed. In this setting, positive MTR gives a prior that a college diploma

has a positive effect on labor market outcome, as mentioned above, this assumption can be explained by both human capital accumulation theory and signaling theory. Positive MTS gives a prior belief that students with characteristics positively correlated with labor market outcomes are more likely to attain a college diploma. These characteristics can be ability, ambitious. Etc. Combining MTR and MTS assumption assumes people with labor-market-preferred characteristics choose to attain a college diploma because they have better knowledge that getting a diploma helps them in job seeking.

Table 3.2 and 3.3 shows the estimation results. Table 3.2 shows treatment effect of college on employment status for unemployed workers in each period. Without other assumptions, the bound in column 1 and column 2 give a wide range. The bound gets larger across years because of decreasing variability. Once positive selection into treatment is assumed, the upper bound becomes very small and shows a decreasing trend. In contrast, when negative selection is assumed, the upper bound shows an increasing trend. There is no apparent pattern in how does the treatment effect evolute. It shows employment status is sometimes not an ideal measure of the labor market outcome. If workers with a college degree have higher reserve wage, they may choose to keep seeking jobs instead of taking currently available positions. Variation in reserve wage gives a wide range of treatment effect on employment status. In the next exercise, income being above the median is used as destination state. When the assumption is that student with higher ability chose to attain a college degree, the treatment effect is very close to zero. Moreover, both upper

bound and lower bound are relatively stable over the years, with upper bound slightly increasing over the years. It reveals one possibility that gaining more human capital is an essential mechanism behind how students with benefit from a college degree. In both Table 3.2 and Table 3.3, there exist periods where treatment effect is not identified under SP assumption, especially under the assumption that positive MTR and positive MTS hold at the same time. Thus this assumption may not be valid in education return research. However, this non-identification happen at late periods; thus it can also be driven by the loss of precision because of the small fraction of survivors.

Table 3.1 shows the difference in the percentile of income. The plus sign represents case where $F_{Y_t^d}(\cdot|\overline{Y}^1_{t-1}, \overline{Y}^0_{t-1}, D = 1)$ first order dominates $F_{Y_t^d}(\cdot|\overline{Y}^1_{t-1}, \overline{Y}^0_{t-1}, D = 0)$. Under this assumption, upper bound to the difference of percentiles is identified. This assumption is equivalent to positive MTR assumption. If we assume the unobserved heterogeneity is ability, the results from Table 3.1 shows that individuals with high ability benefit from college degree through a getting income higher than the 75th percentile. The minus sign represents the reverse, and under this assumption, lower bounds on the difference of percentiles are identified. Also, from the results in Table 3.1, individuals with lower ability tend to benefit from college degree through not falling below 25 percentile in income distribution, which is consistent with results found using binary variables of income being above the 75th percentile and above 25the percentile.

Table 3.1Effect of College Diploma on Income Percentile Differences

Year	25th percentile +	25th percentile -	50th percentile +	50th percentile -	75th percentile +	75th percentile -
1981	$\leq$-150	$\geq$1470	$\leq$240	$\geq$240	$\leq$3521	$\geq$-150
1982	$\leq$0	$\geq$700	$\leq$0	$\geq$0	$\leq$0	$\geq$0
1983	$\leq$0	$\geq$200	$\leq$0	$\geq$0	$\leq$200	$\geq$0
1984	$\leq$0	$\geq$300	$\leq$0	$\geq$0	$\leq$1051	$\geq$0
1985	$\leq$0	$\geq$0	$\leq$0	$\geq$0	$\leq$0	$\geq$0
1986	$\leq$0	$\geq$0	$\leq$0	$\geq$0	$\leq$1000	$\geq$0
1987	$\leq$0	$\geq$160	$\leq$0	$\geq$0	$\leq$160	$\geq$0
1988	$\leq$0	$\geq$480	$\leq$1500	$\geq$1500	$\leq$9000	0
1989	$\leq$0	$\geq$0	$\leq$0	$\geq$0	$\leq$0	$\geq$0
1990	$\leq$400	$\geq$0	$\leq$400	$\geq$400	$\leq$400	$\geq$400

The plus sign "+" stands for cases where $F_{Y_t^d|\overline{Y}^1_{t-1},\overline{Y}^0_{t-1},D=1}$ first order stochastic dominates $F_{Y_t^d\overline{Y}^1_{t-1},\overline{Y}^0_{t-1},D=0}$; the minus sign "-" stands for cases where $F_{Y_t^d|\overline{Y}^1_{t-1},\overline{Y}^0_{t-1},D=0}$ first order stochastic dominates $F_{Y_t^d|\overline{Y}^1_{t-1},\overline{Y}^0_{t-1},D=1}$.

3.5 Discussion

This paper tightens partial identification bounds in a non-parametric dynamic setting based on work of Vickström et al. (2016) and proposes a new method to estimate dynamic treatment effect on treated survivors with non-random treatment assignment. Applying this method to estimate labor market return of college degree, we find college degree has a small signaling effect on employment status and income for individuals with labor-market-preferred characteristics, but potentially high and relatively stable effect for the others, which may come from human capital accumulation. So college diploma affect people with different characteristics through different paths.

It is essential to understand the advantage and disadvantage of partial identification over the other estimation methods. The key advantage is the non-parametric setting. The parametric and semi-parametric setting would induce bias with imposing wrong assumptions on distribution or model structure. Non-parametric estimators are more flexible, giving more accurate information with appropriate assumptions. Partial identification may not be preferred when the range is too general to be informative. However, if sharp bounds can be found, the inference interval of partial identification bounds can be comparable with inference interval in point estimation, and used a test of performance of point estimation. The main shortcoming of this method lies in discrete time assumption. The treatment effects can be sensitive to the way time periods are divided. Also, the assumptions imposed in non-randomization part is not valid in all settings. An extension can be made

to identify the bounds under weaker assumptions. Next step of this paper is to construct a confidence interval so that the results can be comparable with point estimation results.

Table 3.2ATETS of College Diploma on Employment

			MTR				MTR				SP			
			Positive		Negative		Positive		Negative		Positive		Negative	
Year	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB
1981	-0.61	0.64	-0.61	0.13	-0.30	0.64	0.00	0.64	-0.61	0	0	0.13	-0.30	0.00
1982	-0.61	0.78	-0.61	0.10	-0.44	0.78	0.00	0.78	-0.61	0.00	0.00	0.10	-0.44	0
1983	-0.61	0.82	-0.61	0.25	-0.30	0.82	0.00	0.82	-0.61	0.00	0.00	0.25	-0.30	0.00
1984	-0.66	0.80	-0.66	0.13	-0.44	0.80	0.00	0.80	-0.66	0.00	0.00	0.13	-0.44	0.00
1985	-0.67	0.83	-0.67	0.31	-0.27	0.83	0.00	0.83	-0.67	0.00	0.00	0.31	-0.27	0.00
1986	-0.69	0.85	-0.69	0.05	-0.56	0.85	0.00	0.85	-0.69	0.00	0.00	0.05	-0.56	0.00
1987	-0.71	0.82	-0.71	0.32	-0.27	0.82	0.00	0.82	-0.71	0.00	0.00	0.32	-0.27	0.00
1988	-0.73	0.89	-0.73	-0.11	-0.78	0.89	0.00	0.89	-0.73	0.00	N	N	-0.78	0.00
1989	-0.72	0.91	-0.72	-0.09	-0.77	0.91	0.00	0.91	-0.72	0.00	N	N	-0.77	0.00
1990	-0.73	0.87	-0.73	0.12	-0.52	0.87	0.00	0.87	-0.73	0.00	0.00	0.12	-0.52	0.00

The letter "N" represents cases where the bounds can not be identified. This happens under SP assumption when the upper bound is smaller than lower bound.

Table 3.3ATETS of College Diploma on Income Above Media

			MTR				MTR				SP			
			Positive		Negative		Positive		Negative		Positive		Negative	
Year	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB	LB	UB
1981	-0.46	0.75	-0.46	0.08	-0.47	0.75	0.00	0.75	-0.46	0.00	0.00	0.08	-0.47	0.00
1982	-0.47	0.82	-0.47	0.08	-0.50	0.82	0.00	0.82	-0.47	0.00	0.00	0.08	-0.50	0.00
1983	-0.47	0.87	-0.47	0.11	-0.49	0.87	0.00	0.87	-0.47	0.00	0.00	0.11	-0.49	0.00
1984	-0.47	0.90	-0.47	0.16	-0.43	0.90	0.00	0.90	-0.47	0.00	0.00	0.16	-0.43	0.00
1985	-0.47	0.92	-0.47	0.14	-0.46	0.92	0.00	0.92	-0.47	0.00	0.00	0.14	-0.46	0.00
1986	-0.47	0.92	-0.47	0.00	-0.57	0.92	0.00	0.92	-0.47	0.00	0.00	0.00	-0.57	0.00
1987	-0.45	0.94	-0.45	0.10	-0.45	0.94	0.00	0.94	-0.45	0.00	0.00	0.10	-0.45	0.00
1988	-0.47	0.93	-0.47	0.00	-0.56	0.93	0.00	0.93	-0.47	0.00	0.00	0.00	-0.56	0.00
1989	-0.47	0.95	-0.47	0.05	-0.52	0.95	0.00	0.95	-0.47	0.00	0.00	0.05	-0.52	0.00
1990	-0.47	0.92	-0.47	-0.04	-0.56	0.92	0.00	0.92	-0.47	0.00	N	N	-0.56	0.00

Appendices

Appendix A

Appendix for Chapter 1

A.1 Proofs

A.1.1 Proof of Proposition 1.4.1;

$\hat{\epsilon}$ in 1.4.4 is defined as:

$$\begin{aligned}\hat{\epsilon} = \frac{R^Y}{p_y}\left[Y - \left(\frac{R^D}{p_d}g(D,X;\beta) + (1-\frac{R^D}{p_d})E\left[g(D,X;\beta)|Z,X\right]\right)\right] \\ + \left(1-\frac{R^Y}{p_y}\right)\left[\frac{R^D}{p_d}\left(E\left[Y|D,Z,X\right] - g(D,X;\beta)\right)\right. \\ \left. + \left(1-\frac{R^D}{p_d}\right)E\left\{E\left[Y|D,Z,X\right] - g(D,X;\beta)|Z,X\right\}\right]\end{aligned}$$

We expand the right hand side and get

$$
\begin{aligned}
\hat{\epsilon} = & \frac{R^D R^Y}{p_d p_y}\left(Y - g(D, X; \beta)\right) + \left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_d p_y}\right)\left(Y - E\left[g(D, X; \beta)|Z, X\right]\right) \\
& + \left(\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_d p_y}\right)\left(E\left[Y|D, Z, X\right] - g(D, X; \beta)\right) \\
& + \left(1 - \frac{R^D}{p_d}\right)\left(1 - \frac{R^Y}{p_y}\right)\left(E\left[Y|D, Z, X\right] - g(D, X; \beta)\right) \\
= & \frac{R^D R^Y}{p_d p_y}\left(Y - g(D, X; \beta)\right) + \left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_d p_y}\right)\left(Y - E\left[g(D, X; \beta)|Z, X\right]\right) \\
& + \left(\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_d p_y}\right)\left(E\left[Y|D, Z, X\right] - g(D, X; \beta)\right) \\
& + \left[(1 - \frac{R^D R^Y}{p_d p_y}) - \left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_d p_y}\right) - \left(\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_d p_y}\right)\right] \\
& \times \left(E\left[Y|D, Z, X\right] - E\left[g(D, X; \beta)|Z, X\right]\right) \\
= & \frac{R^D R^Y}{p_{11}}\left(Y - g(D, X; \beta)\right) \\
& + \left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right)\left[\left(Y - E\left[g(D, X; \beta)|Z, X\right]\right) - \left(E\left[Y|D, Z, X\right] - E\left[g(D, X; \beta)|Z, X\right]\right)\right] \\
& + \left(\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_{11}}\right)\left[\left(E\left[Y|D, Z, X\right] - g(D, X; \beta)\right) - \left(E\left[Y|D, Z, X\right] - E\left[g(D, X; \beta)|Z, X\right]\right)\right] \\
& + \left(1 - \frac{R^D R^Y}{p_{11}}\right)\left(E\left[Y|D, Z, X\right] - E\left[g(D, X; \beta)|Z, X\right]\right)
\end{aligned}
$$

In the third equality, p_{11} is replaced by $p_d p_y$, because $\frac{R^D R^Y}{p_d p_y} = 0, \forall Z, X$ if $R^D = 0$. Therefore,

$$
\frac{R^D R^Y}{p_d p_y} \equiv \frac{R^D R^Y}{p_d(Z, X) p_y(Z, X, R^D D, R^D)} = \frac{R^D R^Y}{p_d(Z, X) p_y(Z, X, D, R^D = 1)} = \frac{R^D R^Y}{p_{11}} \tag{A.1.1}
$$

A.1.2 Proof of Theorem 1.4.1

We proceed the proof term by term, we first show the expected value of the first term equals to zero:

$$E\left[\frac{R^D R^Y}{p_{11}} Z(Y - g(D, X; \beta))\right] = E\left[Z(Y - g(D, X; \beta))\right] = 0$$

where the first equality sign follows $E\left[\frac{R^D R^Y}{p_{11}} | Z, X, D, Y\right] = 1$, and the second equality sign follows the population moment condition.

Analogously, the zero expectation of the latter three terms can be proved by

$$\begin{aligned}
E\left[\left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right) | Z, X, D, Y\right] &= 0 \\
E\left[\left(\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_{11}}\right) | Z, X, D, Y\right] &= 0 \\
E\left[\left(1 - \frac{R^D R^Y}{p_{11}}\right) | Z, X, D, Y\right] &= 0
\end{aligned}$$

Thus, $E\left[m_{aipw}(\beta)\right] = 0$.

A.1.3 Proof of Theorem 1.5.1

The AIPW estimator maintains the property of double robustness under the assumption MAR. In this section, we give a detailed discussion to show the moment condition holds when either the missing mechanism or the imputed values of missing variables is correctly specified.

First, we show the moment condition specified in Theorem 1.4.1 holds when the missing mechanism is correctly specified. The expectation of moment function is written as:

$$E\left[m_{aipw}(\beta)\right] = E\left[\frac{R^D R^Y}{p_{11}} Z(Y - g(D, X; \beta))\right] + E\left[\phi(Z, X, R^D, R^Y, R^D D, R^Y Y, \beta)\right] \tag{A.1.2}$$

The first term in equation A.1.2 equals to the expectation of full data moment function, and equals to zero:

$$\begin{aligned}
E\left[\frac{R^D R^Y}{p_{11}} Z(Y - g(D, X; \beta))\right] &= E\left\{E\left[\frac{R^D R^Y}{p_{11}} Z(Y - g(D, X; \beta))|Z, X, D, Y\right]\right\} \\
&= E\left\{E\left[\frac{R^D R^Y}{p_{11}}|Z, X, D, Y\right] Z(Y - g(D, X; \beta))\right\} \\
&= E\left\{\frac{\Pr\left[R^D = 1, R^Y = 1|Z, X, D, Y\right]}{p_{11}} Z(Y - g(D, X; \beta))\right\} \\
&= E\left[Z(Y - g(D, X; \beta))\right] = 0
\end{aligned}$$

The last equality follows the full data moment condition defined in equation 1.2.4.

The second term in equation A.1.2 equals to zero following the analogous argument. We show the expectation of each element in ϕ equals to zero:

$$
\begin{aligned}
&E\left\{\left(\frac{R^Y}{p_y}-\frac{R^DR^Y}{p_{11}}\right)Z\left[(Y-E\left[g(D,X;\beta)|Z,X\right])-(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right])\right]\right\}\\
&=E\left\{E\left[\left(\frac{R^Y}{p_y}-\frac{R^DR^Y}{p_{11}}\right)a(Z,X,Y)|Z,X,D,Y\right]\right\}\\
&=E\left\{E\left[\left(\frac{R^Y}{p_y}-\frac{R^DR^Y}{p_{11}}\right)|Z,X,D,Y\right]a(Z,X,Y)\right\}\\
&=E\left\{\left(\frac{\Pr\left[R^Y=1|Z,X,D,Y\right]}{p_y}-\frac{\Pr\left[R^D=1,R^Y=1|Z,X,D,Y\right]}{p_{11}}\right)a(Z,X,Y)\right\}\\
&=0
\end{aligned}
$$

where a is a function of the observables such that

$$
a(Z,X,Y)=Z\left[(Y-E\left[g(D,X;\beta)|Z,X,Y\right])-(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right])\right]
$$

The last equality holds if p_{11}, p_y are correctly specified.

The second term follows the same argument.

$$
\begin{aligned}
&E\left[\left(\frac{R^D}{p_d}-\frac{R^DR^Y}{p_{11}}\right)|Z,X,D,Y\right]\\
&=\frac{\Pr\left[R^D=1|Z,X,D,Y\right]}{p_d}-\frac{\Pr\left[R^D=1,R^Y=1|Z,X,D,Y\right]}{p_{11}}=0
\end{aligned}
$$

if p_d is also correctly specified.

And the third term equals to 0, following

$$
E\left[\left(1-\frac{R^DR^Y}{p_{11}}\right)|Z,X,D,Y\right]=1-\frac{\Pr\left[R^D=1,R^Y=1|Z,X,D,Y\right]}{p_{11}}=0
$$

Therefore, the moment condition holds if (p_d, p_y, p_{11}) are correctly specified.

Next, we show the moment condition holds if $E[g(D,X;\beta)|Z,X]$, $E[Y|D,Z,X]$, $E[Y|Z,X]$ are correctly specified. We first consider the first component of the augmentation term ϕ:

$$
\begin{aligned}
&E\left\{\left(\frac{R^Y}{p_{11}+p_{01}}-\frac{R^D R^Y}{p_{11}}\right) Z\left[(Y-E[g(D,X;\beta)|Z,X,Y])-(E[Y|Z,X]-E[g(D,X;\beta)|Z,X])\right]\right. \\
&= E\left\{\left(\frac{R^Y}{p_{11}+p_{01}}-\frac{R^D R^Y}{p_{11}}\right) E[a(Z,X,Y)|Z,X]\right\} \\
&= 0
\end{aligned}
$$

following $E[Y-E[g(D,X;\beta)|Z,X,Y]|Z,X] = E[Y|Z,X] - E[g(D,X;\beta)|Z,X]$.

The same proof can be applied on the second component of ϕ, following

$$E[E[Y|D,Z,X]-g(D,X;\beta)|Z,X] = E[Y|Z,X]-E[g(D,X;\beta)|Z,X]$$

Next, we combine the last term in ϕ and the IPW moment condition and get:

$$
\begin{aligned}
&E\left\{\frac{R^D R^Y}{p_{11}}[(Y-g(D,X;\beta))-(E[Y|Z,X]-E[g(D,X;\beta)|Z,X])]\right. \\
&+(E[Y|Z,X]-E[g(D,X;\beta)|Z,X])\} \\
&= 0 + E\{E[Y|Z,X]-E[g(D,X;\beta)|Z,X]\} \\
&= E[m_{aipw}(\beta)] = 0
\end{aligned}
$$

The second equality follows:

$$E\left[Y - g(D, X; \beta)|Z, X\right] = E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right]$$

and the last equality follows the iterative law of expectations.

A.1.4 Proof of Theorem 1.5.2

We use m_{full} to denote the original population level moment function, such that:

$$m_{full}(\beta) = Z\left(Y - g(D, X; \beta)\right)$$

We reorganize the moment function into:

$$\begin{aligned} m_{aipw}(\beta) = & \underbrace{\frac{R^D R^Y}{p_{11}}\left(m_{full} - E\left[m_{full}|Z, X\right]\right)}_{(1)} + \underbrace{E\left[m_{full}|Z, X\right]}_{(2)} \\ & + \underbrace{\left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right)\left(E\left[m_{full}|D, Z, X\right] - E\left[m_{full}|Z, X\right]\right)}_{(3)} \\ & + \underbrace{\left(\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_{11}}\right)\left(Y - E\left[Y|Z, X\right]\right)}_{(4)} \end{aligned}$$

We discuss the variance of each term.

The variance of the first term equals to:

$$
\begin{aligned}
Var\left((1)\right) &= E\left[Var\left(\frac{R^D R^Y}{p_{11}}\left(m_{full} - E\left[m_{full}|Z,X\right]\right)|Z,X\right)\right] \\
&\quad + Var\left(E\left[\frac{R^D R^Y}{p_{11}}\left(m_{full} - E\left[m_{full}|Z,X\right]\right)|Z,X\right]\right) \\
&= E\left[\frac{1}{p_{11}} Var\left(\left(m_{full} - E\left[m_{full}|Z,X\right]\right)|Z,X\right)\right] = E\left[\frac{1}{p_{11}} Var\left(m_{full}|Z,X\right)\right]
\end{aligned}
$$

The variance of the second term equals to:

$$
Var\left((2)\right) = E\left[E\left[m_{full}|Z,X\right]E\left[m_{full}|Z,X\right]'\right]
$$

The covariance between (1) and (2) equals to 0 following:

$$
\begin{aligned}
Cov\left((1),(2)\right) &= E\left[\frac{R^D R^Y}{p_{11}}\left(m_{full} - E\left[m_{full}|Z,X\right]\right)E\left[m_{full}|Z,X\right]\right] \\
&= E\left[\left(E\left[m_{full}|Z,X\right] - E\left[m_{full}|Z,X\right]\right)E\left[m_{full}|Z,X\right]\right] = 0
\end{aligned}
$$

The covariance between (1) and (3) is equivalent to the negative of $Var\left((3)\right)$, following

$$
\begin{aligned}
Cov\left((1),(3)\right) &= Cov\left((1), \frac{R^D}{p_d}\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)\right) \\
&\quad - Cov\left((1), \frac{R^D R^Y}{p_{11}}\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)\right) \\
&= E\left[\frac{1}{p_d}\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)^2\right] \\
&\quad - E\left[\frac{1}{p_{11}}\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)^2\right] \\
&= -Var\left((3)\right)
\end{aligned}
$$

Also, $Cov((2),(3)) = 0$ following similar argument as in the proof for $Cov\left((1),(2)\right)$.

Next, we derive the correlation between (3) and the other terms.

$$
\begin{aligned}
Cov\left((1),(4)\right) &= E\left[\frac{R^D R^Y}{p_{11}p_y}\left(m_{full} - E\left[m_{full}|Z,X\right]\right)Z\left(Y - E\left[Y|Z,X\right]\right)\right] \\
&\quad - E\left[\frac{R^D R^Y}{p_{11}^2}\left(m_{full} - E\left[m_{full}|Z,X\right]\right)Z\left(Y - E\left[Y|Z,X\right]\right)\right] \\
&= E\left[\left(\frac{1}{p_y} - \frac{1}{p_{11}}\right)E\left[\left(m_{full} - E\left[m_{full}|Z,X\right]\right)Z\left(Y - E\left[Y|Z,X\right]\right)|Z,X\right]\right] \\
&= -Var\left((4)\right) - E\left[\left(\frac{1}{p_y} - \frac{1}{p_{11}}\right)Z^2 Cov\left(g(D,X;\beta), Y|Z,X\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
Cov\left((3),(4)\right) &= E\left[\left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right)\left(\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_{11}}\right)\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)\right. \\
&\quad \left. Z\left(Y - E\left[Y|Z,X\right]\right)\right] \\
&= E\left[\left(1 - \frac{1}{p_y} - \frac{1}{p_d} + \frac{1}{p_{11}}\right)Cov\left(ZE\left[Y|D,Z,X\right], E\left[m_{full}|D,Z,X\right]|Z,X\right)\right]
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
V_{MAR} = & E\left[\frac{1}{p_{11}} Var\left(m_{full}|Z,X\right) + E\left[m_{full}|Z,X\right] E\left[m_{full}|Z,X\right]'\right] \\
& - Var\left(\left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right)\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)\right) \\
& - Var\left(\left(\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_{11}}\right)\left(E\left[m_{full}|Y,Z,X\right] - E\left[m_{full}|Z,X\right]\right)\right) \\
& - 2E\left[\left(1 - \frac{1}{p_y} - \frac{1}{p_d} + \frac{1}{p_{11}}\right) Cov\left(ZE\left[Y|D,Z,X\right], E\left[m_{full}|D,Z,X\right]|Z,X\right)\right] \\
& - 2E\left[\left(\frac{1}{p_y} - \frac{1}{p_{11}}\right) Z^2 Cov\left(g(D,X;\beta), Y|Z,X\right)\right]
\end{aligned}
$$

A.1.5 Proof of Theorem 1.5.4

The first part of this proof heavily depend on the results from Newey (1994) and Newey and McFadden (1994). Two takeaways from Newey (1994) are: (a) the methods of estimating the nuisance parameters do not affect the asymptotic variance of the estimator; (b) the nuisance parameters do not affect the variance of the primary model if it does not affect consistency of the model.

First, we rewrite the moment function as:

$$
\begin{aligned}
m_{aipw}(\beta) = & \frac{R^D R^Y}{p_{11}} Z(Y - g(D, X; \beta)) \\
& + (1 - p_d) \left(\frac{(1 - R^D) R^Y}{p_{01}} - \frac{R^D R^Y}{p_{11}} \right) \\
& \times Z \left[(Y - E \left[g(D, X; \beta) | Z, X \right]) - (E \left[Y | Z, X \right] - E \left[g(D, X; \beta) | Z, X \right]) \right] \\
& + \left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}} \right) Z \left[(E \left[Y | D, Z, X \right] - g(D, X; \beta)) \right. \\
& \left. - (E \left[Y | Z, X \right] - E \left[g(D, X; \beta) | Z, X \right]) \right] \\
& + \left(1 - \frac{R^D R^Y}{p_{11}} \right) Z \left(E \left[Y | Z, X \right] - E \left[g(D, X; \beta) | Z, X \right] \right)
\end{aligned}
$$

by replacing

$$
\begin{aligned}
\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_{11}} &= \frac{R^D R^Y}{p_y} + \frac{(1 - R^D) R^Y}{p_y} - \frac{R^D R^Y}{p_{11}} \\
&= \frac{p_d R^D R^Y}{p_{11}} + \frac{(1 - p_d)(1 - R^D) R^Y}{p_{01}} - \frac{R^D R^Y}{p_{11}} \\
&= (1 - p_d) \left(\frac{(1 - R^D) R^Y}{p_{01}} - \frac{R^D R^Y}{p_{11}} \right)
\end{aligned}
$$

in the second component in m_{aipw}.

The nuisance parameters include the probability of observing D, Y, and both, conditional on the observables. Recall that the imputed values for incomplete model, i.e, $E \left[g(D, X; \beta) | Z, X \right]$, $E \left[Y | Z, X \right]$, $E \left[Y | D, Z, X \right]$ do not affect the moment condition, they do not have effect on the variance of V_{SMAR}. On the other hand, among the propensities, only p_{11} affect consistency of m_{aipw} via the first component, therefore, despite existence of multiple nuisance

parameters in four components, we need to only create the correction term for one of them. For simplicity of notation, denote

$$\begin{aligned}\phi_c &= -(1-p_d)\frac{R^D R^Y}{p_{11}} Z\left[(Y - E\left[g(D,X;\beta)|Z,X\right]) - (E\left[Y|Z,X\right] - E\left[g(D,X;\beta)|Z,X\right])\right] \\ &= -(1-p_d)\frac{R^D R^Y}{p_{11}} Z\left[Y - E\left[Y|Z,X\right]\right]\end{aligned}$$

Derivation of the correction term closely followsNewey (1994). Note that p_{11} can be seen as a conditional expectation of $R^D R^Y$. Results from Section 4 in Newey (1994) can be applied. We first need to find a linearization of ϕ_c. Because p_{11} only affects ϕ_c through its values instead of functional form, we apply equation (3.17) from Newey (1994) directly, and derive the linearization of ϕ_c as:

$$D(\mathcal{O}, \phi_c) = E\left[(1-p_d)\frac{1}{p_{11}}\left(E\left[Y|D,Z,X\right] - E\left[Y|Z,X\right]\right)\hat{p}_{11}(D,Z,X)\right]$$

and from result in (4.5),

$$\begin{aligned}\frac{\partial E\left[D(\mathcal{O},\phi_c)\right]}{\partial \theta} &= E\left\{(1-p_d)\frac{1}{p_{11}}\left(E\left[Y|D,Z,X\right] - E\left[Y|Z,X\right]\right)\left(R^D R^Y - p_{11}\right)\mathcal{S}(\mathcal{O})\right\} \\ &= E\left[(1-p_d)(\frac{R^D R^Y}{p_{11}} - 1)\left(E\left[Y|D,Z,X\right] - E\left[Y|Z,X\right]\right)\mathcal{S}(\mathcal{O})\right]\end{aligned}$$

Therefore, the correction term to be incorporated is

$$(1-p_d)(1 - \frac{R^D R^Y}{p_{11}})\left(E\left[Y|D,Z,X\right] - E\left[Y|Z,X\right]\right)$$

Now, we proceed calculation of the V_{SMAR} term by term. We reorganize the influence function as in Section A.1.4:

$$\begin{aligned}\phi_{SMAR}(\beta) = & \underbrace{\frac{R^D R^Y}{p_{11}}\left(m_{full} - E\left[m_{full}|Z,X\right]\right)}_{(1)} + \underbrace{E\left[m_{full}|Z,X\right]}_{(2)} \\ & + \underbrace{\left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right)\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)}_{(3)} \\ & + \underbrace{\left(\frac{R^Y}{p_y} - \frac{R^D R^Y}{p_{11}}\right)\left(Y - E\left[Y|Z,X\right]\right)}_{(4)} \\ & + \underbrace{(1-p_d)(\frac{R^D R^Y}{p_{11}} - 1)\left(E\left[Y|D,Z,X\right] - E\left[Y|Z,X\right]\right)}_{(5)}\end{aligned}$$

where (5) is the correction term added for nuisance parameter used in (3).

The variance of the first term equals to:

$$
\begin{aligned}
Var\left((1)\right) &= E\left[Var\left(\frac{R^D R^Y}{p_{11}}\left(m_{full} - E\left[m_{full}|Z,X\right]\right)|D,Z,X\right)\right] \\
&\quad + Var\left(E\left[\frac{R^D R^Y}{p_{11}}\left(m_{full} - E\left[m_{full}|Z,X\right]\right)|D,Z,X\right]\right) \\
&= E\left[\frac{1}{p_{11}} Var\left(\left(m_{full} - E\left[m_{full}|Z,X\right]\right)|D,Z,X\right)\right] \\
&\quad + Var\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right) \\
&= E\left[\frac{1}{p_{11}} Var\left(\left(m_{full} - E\left[m_{full}|Z,X\right]\right)|D,Z,X\right)\right] + Var\left(E\left[m_{full}|D,Z,X\right]\right) \\
&\quad + Var\left(E\left[m_{full}|Z,X\right]\right) - 2Cov\left(E\left[m_{full}|D,Z,X\right], E\left[m_{full}|Z,X\right]\right) \\
&= E\left[\frac{1}{p_{11}} Var\left(m_{full}|D,Z,X\right)\right] + Var\left(E\left[m_{full}|D,Z,X\right]\right) - Var\left(E\left[m_{full}|Z,X\right]\right) \\
&= E\left[\frac{1}{p_{11}} Var\left(m_{full}|D,Z,X\right)\right] + Var\left(E\left[m_{full}|D,Z,X\right]\right) - Var\left((2)\right)
\end{aligned}
$$

The $-Var((2))$ cancels out with the variance of the second term.

Now we consider the covariance between different terms:

$$Cov\left((1),(2)\right) = 0, Cov\left((2),(3)\right) = 0, Cov\left((2),(4)\right) = 0$$

$Cov\left((2),(5)\right)$ also equals to zero following analogous argument and the fact that $p_{11} = E\left[R^D R^Y|Z,X,D,Y\right]$.

$$Cov\left((1),(3)\right) = -Var((3))$$

following analogous argument in the Section A.1.4.

Now we consider the covariance between (1) and (4):

$$
\begin{aligned}
Cov\left((1),(4)\right) &= E\left[-(1-p_d)\frac{R^D R^Y}{p_{11}^2}\left(m_{full} - E\left[m_{full}|Z,X\right]\right)\left(Y - E\left[Y|Z,X\right]\right)\right] \\
&= E\left[-\frac{1-p_d}{p_{11}} E\left[\left(m_{full} - E\left[m_{full}|Z,X\right]\right)\left(Y - E\left[Y|Z,X\right]\right)|D,Z,X\right]\right]
\end{aligned}
$$

We notice that the covariance between (1) and (5) equals to:

$$\begin{aligned} Cov\left((1),(5)\right) = E\Bigg[&(1-p_d)\left(\frac{R^D R^Y}{p_{11}^2} - \frac{R^D R^Y}{p_{11}}\right)\left(m_{full} - E\left[m_{full}|Z,X\right]\right) \\ &\times \left(E\left[Y|D,Z,X\right] - E\left[Y|Z,X\right]\right)\Bigg] \\ = E\Bigg[&(1-p_d)\left(\frac{1}{p_{11}} - 1\right)\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right) \\ &\times \left(E\left[Y|D,Z,X\right] - E\left[Y|Z,X\right]\right)\Bigg] \end{aligned}$$

Therefore, some parts in $Cov\left((1),(4)\right)$ and $Cov\left((1),(5)\right)$ cancels with each other, and get:

$$\begin{aligned} Cov\left((1),(4)\right) + Cov\left((1),(5)\right) = &-E\left[\frac{1-p_d}{p_{11}} Var\left(Y|D,Z,X\right)\right] \\ &- E\left[(1-p_d) Cov\left(E\left[m_{full}|D,Z,X\right], E\left[Y|D,Z,X\right]|Z,X\right)\right] \end{aligned}$$

The variance of (4) equals to:

$$Var\left((4)\right) = E\left[(1-p_d)^2\left(\frac{1}{p_{01}} + \frac{1}{p_{11}}\right) Var\left(Y|D,Z,X\right)\right]$$

Next, consider covariance between $(3),(4),(5)$

$$
\begin{aligned}
&Cov\left((3),(4)\right) \\
&= E\left[-(1-p_d)\left(\frac{R^D R^Y}{p_d p_{11}} - \frac{R^D R^Y}{p_{11}^2}\right)\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)\left(Y - E\left[Y|Z,X\right]\right)\right] \\
&= E\left[-(1-p_d)\left(\frac{1}{p_d} - 1\right)\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)\left(E\left[Y|D,Z,X\right] - E\left[Y|Z,X\right]\right)\right] \\
&= E\left[-\frac{(1-p_d)^2}{p_d} Cov\left(E\left[m_{full}|D,Z,X\right], E\left[Y|D,Z,X\right]|Z,X\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&Cov\left((3),(5)\right) \\
&= E\left[(1-p_d)\left(\frac{R^D R^Y}{p_d p_{11}} - \frac{R^D R^Y}{p_{11}^2}\right)\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)\left(Y - E\left[Y|Z,X\right]\right)\right] \\
&\quad - E\left[(1-p_d)\left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right)\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)\left(Y - E\left[Y|Z,X\right]\right)\right] \\
&= E\left[\frac{(1-p_d)^2}{p_d} Cov\left(E\left[m_{full}|D,Z,X\right], E\left[Y|D,Z,X\right]|Z,X\right)\right]
\end{aligned}
$$

The two covariances cancel with each other.

$$
\begin{aligned}
&Cov\left((4),(5)\right) \\
&= E\left[-(1-p_d)^2\left(\frac{R^D R^Y}{p_{11}^2} - \frac{R^D R^Y}{p_{11}}\right)\left(E\left[Y|D,Z,X\right] - E\left[Y|Z,X\right]\right)\left(Y - E\left[Y|Z,X\right]\right)\right] \\
&= E\left[(1-p_d)^2\left(1 - \frac{1}{p_{11}}\right) Var\left(E\left[Y|D,Z,X\right]|Z,X\right)\right] \\
&= -Var\left((5)\right)
\end{aligned}
$$

Thus, V_{SMAR} is derived as follows:

$$
\begin{aligned}
V_{SMAR} = & E\left[\frac{1}{p_{11}} Var\left(\left(m_{full} - E\left[m_{full}|Z,X\right]\right)|D,Z,X\right)\right] + Var\left(E\left[m_{full}|D,Z,X\right]\right) \\
& - Var\left(\left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right)\left(E\left[m_{full}|D,Z,X\right] - E\left[m_{full}|Z,X\right]\right)\right) \\
& - Var\left((1-p_d)(\frac{R^D R^Y}{p_{11}} - 1)\left(E\left[Y|D,Z,X\right] - E\left[Y|Z,X\right]\right)\right) \\
& - E\left[\left(\frac{1-p_d^2}{p_{11}} - \frac{(1-p_d)^2}{p_{01}}\right) Var\left(Y|D,Z,X\right)\right] \\
& - 2E\left[(1-p_d) Cov\left(E\left[m_{full}|D,Z,X\right], E\left[Y|D,Z,X\right]|Z,X\right)\right]
\end{aligned}
$$

where the two terms come from $Var\left((4)\right)+2Cov\left((1),(4)\right)+2Cov\left((1),(5)\right)$.

A.1.6 Proof of Theorem 1.5.5

The sketch of the proof follows Cattaneo (2010), Chaudhuri and Guilkey (2016), Newey (1994), Whitney K Newey (1997) closely, and most of the results have been proved in the proof of Proposition 2.3 in Chaudhuri and Guilkey (2016).

First, by the results showed in Theorem 5 and Theorem 8 in Cattaneo (2010), under the conditions listed in Theorem 1.5.5,

$$
||\hat{p} - p||_\infty = o_p(N^{-\frac{1}{4}}) \tag{A.1.3}
$$

$$
||\hat{q} - q||_\infty = o_p(N^{-\frac{1}{4}}) \tag{A.1.4}
$$

where p stands for the missing mechanism parameters (p_d, p_y, p_{11}), and q stands for the imputed missing values $\left(E\left[D|Z,X\right], E\left[Y|Z,X\right], E\left[Y|Z,X,D\right]\right)$.

Then, what left is to show the multiplication of the nuisance parameters converges to zero in probability. We show the sketch for

$$\left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right)\left[\left(E\left[Y|D,Z,X\right] - g(D,X;\beta)\right) - \left(E\left[Y|Z,X\right] - E\left[g(D,X;\beta)|Z,X\right]\right)\right]$$

and we use analogous conditions to the conditions (32) and (33) from the proof of proposition 2.3 in Chaudhuri and Guilkey (2016), what left to show is that

$$o_p(1) = \sqrt{N}\left(\bar{\xi}_N(\beta^0, \hat{p}, \hat{q}(\beta^0)) - \bar{\xi}_N(\beta^0, p, q(\beta^0))\right) \tag{A.1.5}$$

$$o_p(1) = \sup_{|\beta-\beta_0|\leq\delta_N} \frac{\sqrt{N}|\bar{\xi}_N\left(\beta, \hat{p}, \hat{q}(\beta)\right) - E\left[\bar{\xi}_N(\beta, p, q(\beta))\right] - \bar{\xi}_N\left(\beta^0, \hat{p}, \hat{q}(\beta^0)\right)|}{1 + C\sqrt{N}|\beta - \beta^0|} \tag{A.1.6}$$

for all positive sequences $\delta_N = o(1)$ and a generic constant $C > 0$, and

$$\bar{\xi}_N(\beta, \hat{p}, \hat{q}) = \frac{1}{N}\sum_{i=1}^{N} \hat{\nu}_i \hat{\tau}_i$$

$$\bar{\xi}_N(\beta, p\hat{q}) = \frac{1}{N}\sum_{i=1}^{N} \nu_i \tau_i$$

where

$$\nu_i = \left(\frac{R_i^D}{p_{d,i}} - \frac{R_i^D R_i^Y}{p_{11,i}}\right)$$

$$\tau_i = \left[\left(E\left[Y_i|D_i,Z_i,X_i\right] - g(D_i,X_i;\beta)\right) - \left(E\left[Y_i|Z_i,X_i\right] - E\left[g(D_i,X_i;\beta)|Z_i,X_i\right]\right)\right]$$

Condition A.1.5 holds follows the fact that $\frac{1}{N}\sum_{i=1}^{N}|\nu_i|$ and $\frac{1}{N}\sum_{i=1}^{N}|\tau_i|$ are $O_p(1)$, which follows the conditions A.1.3 and A.1.4.

The proof of the condition A.1.6 also follows the proof of the analogous condition in Chaudhuri and Guilkey (2016). First, $E\left[\bar{\xi}_N(\beta, p, q(\beta))\right] = 0$ under either the Assumption MAR or the Assumption SMAR, therefore, the condition A.1.6 reduces to:

$$\begin{aligned} o_p(1) &= \sup_{|\beta-\beta_0|\leq\delta_N} \frac{\sqrt{N}|\bar{\xi}_N\left(\beta, \hat{p}, \hat{q}(\beta)\right) - \bar{\xi}_N\left(\beta^0, \hat{p}, \hat{q}(\beta^0)\right)|}{1 + C\sqrt{N}|\beta - \beta^0|} \\ &= \sup_{|\beta-\beta_0|\leq\delta_N} \frac{|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\hat{\nu}_i\left(\hat{\tau}_i(\beta) - \hat{\tau}_i(\beta^0)\right)|}{1 + C\sqrt{N}|\beta - \beta^0|} \end{aligned} \tag{A.1.7}$$

Proof of equation A.1.7 can be found in proof of Proposition 2.3 and Proposition 2.4 in Chaudhuri and Guilkey (2016), by setting $\omega_i = c$ for some constant c in their corresponding condition. The key condition used in the proof is that $E\left[\tau_i(\beta)\right] = 0$ for any β, and it holds from the definition of $\tau_i(\beta)$.

The analogous proof sketch can be used on the component for $R^Y = 1$, i.e.,

$$\left(\frac{R^D}{p_d} - \frac{R^D R^Y}{p_{11}}\right)\left[\left(E\left[Y|D, Z, X\right] - g(D, X; \beta)\right) - \left(E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right]\right)\right]$$

Next, we combine the IPW moment condition and the last component in the augmentation term into:

$$\frac{R^D R^Y}{p_{11}}\left[(Y - g(D, X; \beta)) - (E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right])\right] + \quad \text{(A.1.8)}$$

$$(E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right]) \quad \text{(A.1.9)}$$

$$\frac{R^D R^Y}{p_{11}}\left[(Y - g(D, X; \beta)) - (E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right])\right] + \quad \text{(A.1.10)}$$

$$(E\left[Y|Z, X\right] - E\left[g(D, X; \beta)|Z, X\right]) \quad \text{(A.1.11)}$$

and proof of convergence of equation A.1.11 can be found in Theorem 8 in Cattaneo (2010).

A.1.7 Proof of Lemma 1

Suppose Assumption SMAR and Overlap hold, and $Y \perp\!\!\!\perp D|Z, X$, then the AIPW estimator maintains double-robustness, following the fact

$$E\left[Y|Z, X\right] = E\left[Y|D, Z, X\right]$$

and therefore estimation on p_{11} does not affect consistency of the primary estimator. Then, there is no correction term needed for the AIPW estimator, under the SMAR assumption. Or, the correction term derived in the section A.1.5 equals to 0 following:

$$\begin{aligned}
&(1 - p_d)(1 - \frac{R^D R^Y}{p_{11}})\left(E\left[Y|D, Z, X\right] - E\left[Y|Z, X\right]\right) \\
&= (1 - p_d)(1 - \frac{R^D R^Y}{p_{11}})\left(E\left[Y|D, Z, X\right] - E\left[Y|D, Z, X\right]\right) = 0
\end{aligned}$$

Therefore, $V_{SMAR} = V_{MAR}$.

A.1.8 Proof of Theorem 1.5.6

We proceed the proof the efficiency following the classical three steps.

Step 1 We denote the fully observed variables as $O = (Z, X, R^D, R^Y, R^D D, R^Y Y)$, and consider a class of parametric submodels indexed by θ such that the distribution of O can be expressed as:

$$\begin{aligned} f_\theta(O) = &\left[p_{\theta,d}(Z,X)p^1_{\theta,y}(Z,X,D)f_\theta(Y|Z,X,D)f_\theta(D|Z,X)\right]^{R^D R^Y} \\ &\times \left[p_{\theta,d}(Z,X)\left(1-p^1_{\theta,y}(Z,X,D)\right)f_\theta(D|Z,X)\right]^{R^D(1-R^Y)} \\ &\times \left[\left(1-p_{\theta,d}(Z,X)\right)p^0_{\theta,y}(Z,X)f_\theta(Y|Z,X)\right]^{(1-R^D)R^Y} \\ &\times \left[\left(1-p^0_{\theta,d}(Z,X)\right)\left(1-p_{\theta,y}(Z,X)\right)\right]^{(1-R^D)(1-R^Y)} f_\theta(Z,X) \end{aligned}$$

where $p^1_{\theta,y}(D,Z,X)$ is defined as the probability of observing Y given $R^D = 1$; and $p^0_{\theta,d}(Z,X)$ defined as the probability of observing Y given $R^D = 0$. They are defined formally as:

$$\begin{aligned} p^1_{\theta,y}(D,Z,X) &= \Pr\left[R^Y = 1|Z,X,D,R^D = 1\right] \\ p^0_{\theta,y}(Z,X) &= \Pr\left[R^Y = 1|Z,X,R^D = 0\right] \end{aligned}$$

The score function is defined as:

$$
\begin{aligned}
\mathcal{S}_\theta(O) &= s_\theta(Z,X) + R^D R^Y s_\theta(Y,D|Z,X) + R^D(1-R^Y)s_\theta(D|Z,X) + (1-R^D)R^Y s_\theta(Y|Z,X) \\
&+ \left\{ R^D R^Y \left(\frac{\dot{p}_{\theta,d}(Z,X)}{p_{\theta,d}(Z,X)} + \frac{\dot{p}^1_{\theta,y}(Z,X,D)}{p^1_{\theta,y}(Z,X,D)} \right) + R^D(1-R^Y) \left(\frac{\dot{p}_{\theta,d}(Z,X)}{p_{\theta,d}(Z,X)} - \frac{\dot{p}^1_{\theta,y}(Z,X,D)}{1-p^1_{\theta,y}(Z,X,D)} \right) \right. \\
&+ (1-R^D)R^Y \left(\frac{\dot{p}^0_{\theta,y}(Z,X)}{p^0_{\theta,y}(Z,X)} - \frac{\dot{p}_{\theta,d}(Z,X)}{1-p_{\theta,d}(Z,X)} \right) \\
&\left. +(1-R^D)(1-R^Y) \left(-\frac{\dot{p}^0_{\theta,y}(Z,X)}{1-p^0_{\theta,y}(Z,X)} - \frac{\dot{p}_{\theta,d}(Z,X)}{1-p_{\theta,d}(Z,X)} \right) \right\} \\
&= s_\theta(Z,X) + R^D R^Y s_\theta(Y,D|Z,X) + R^D(1-R^Y)s_\theta(D|Z,X) + (1-R^D)R^Y s_\theta(Y|Z,X) \\
&+ \left\{ \frac{R^D - p_{\theta,d}(Z,X)}{p_{\theta,d}(Z,X)(1-p_{\theta,d}(Z,X))} \dot{p}_{\theta,d}(Z,X) + \right. \\
&R^D \frac{R^Y - p^1_{\theta,y}(D,Z,X)}{p^1_{\theta,y}(D,Z,X)(1-p^1_{\theta,y}(D,Z,X))} \dot{p}^1_{\theta,y}(D,Z,X) \\
&\left. (1-R^D) \frac{R^Y - p^0_{\theta,y}(Z,X)}{p^0_{\theta,y}(D,Z,X)(1-p^0_{\theta,y}(Z,X))} \dot{p}^0_{\theta,y}(Z,X) \right\}
\end{aligned}
$$

where $s_\theta(Z,X) \equiv \frac{\partial}{\partial\theta} f_\theta(Z,X)$, $s_\theta(Y,D|Z,X) \equiv \frac{\partial}{\partial\theta} f_\theta(Y,D|Z,X)$, $s_\theta(D|Z,X) \equiv \frac{\partial}{\partial\theta} f_\theta(D|Z,X)$, $s_\theta(Y|Z,X) \equiv \frac{\partial}{\partial\theta} f_\theta(Y|Z,X)$, and $\dot{p}_{\theta,d}(Z,X) = \frac{\partial}{\partial\theta} p_{\theta,d}(Z,X)$, $\dot{p}^1_{\theta,y}(Z,X,D) = \frac{\partial}{\partial\theta} p^1_{\theta,y}(Z,X,D)$, $\dot{p}^0_{\theta,y}(Z,X) = \frac{\partial}{\partial\theta} p^0_{\theta,y}(Z,X)$.

The tangent set is characterized by:

$$
\begin{aligned}
\mathcal{T} \equiv\ & R^D R^Y f_{11}(Y,D,Z,X) + R^D(1-R^Y) f_{10}(D,Z,X) + (1-R^D)R^Y f_{01}(Y,Z,X) + f_0(Z,X) \\
&+ R^D R^Y \frac{b_{11}(D,Z,X)}{c_{11}(D,Z,X)} + R^D(1-R^Y)\frac{b_{10}(D,Z,X)}{c_{10}(D,Z,X)} + (1-R^D)R^Y \frac{b_{01}(D,Z,X)}{c_{01}(D,Z,X)} \\
&+ (1-R^D)(1-R^Y)\frac{b_{00}(Z,X)}{c_{00}(Z,X)}
\end{aligned}
$$

where

$$f_{11}(Y,D,Z,X) \in L_0^2\left(F\left(Y,D|Z,X\right)\right), f_{10}(D,Z,X) \in L_0^2\left(F(D|Z,X)\right)$$
$$f_{01}(Y,Z,X) \in L_0^2\left(F\left(Y|Z,X\right)\right), f_0(Z,X) \in L_0^2\left(F(Z,X)\right)$$

Step 2 The full data moment condition is written as:

$$AE\left[m_{full}(Z,X,D,Y;\beta^0)\right] = 0$$

for any matrix A of size $d_\theta \times d_m$, where d_θ is the dimension of unknown parameters while d_m is the number of moment conditions. The matrix A is added to convert an over-identified system of moment conditions into the just-identified moment conditions. In our framework, the moment condition m_{full} is defined to be $m_{full}(Z,X,D,Y;\beta^0) \equiv Z\left(Y - g(D,X;\beta^0)\right)$.

$$\begin{aligned}\frac{\partial}{\partial\theta}\beta^0(\theta_0) &= -(AG)^{-1}AE\left[m_{full}(Z,X,D,Y;\beta^0)\frac{\partial}{\partial\theta}log f_{\theta_0}(Z,X,D,Y)\right] \\ &= -(AG)^{-1}AE\left[m_{full}(Z,X,D,Y;\beta^0)\left(s(Z,X)' + s(D|Z,X)' + s(Y|D,Z,X)'\right)\right]\end{aligned}$$

Then, we conjecture a such that

$$E\left[\varphi\mathcal{S}_O'\right] = E\left[m_{full}(Z,X,D,Y;\beta^0)\left(s(Z,X)' + s(D|Z,X)' + s(Y|D,Z,X)'\right)\right]$$

Then, we confirm term by term, and the first term equals

$$
\begin{aligned}
&E\left[\frac{R^D R^Y}{p_{11}} Z(Y - g(D, X; \beta))\mathcal{S}_O\right] \\
&= E\left\{\frac{R^D R^Y}{p_{11}} Z\left(Y - g(D, X; \beta)\right) \times \left(s_\theta(Z, X) + s_\theta(Y, D|Z, X) + \right.\right. \\
&\quad \frac{R^D - p_{\theta,d}(Z, X)}{p_{\theta,d}(Z, X)(1 - p_{\theta,d}(Z, X))} \dot{p}_{\theta,d}(Z, X) \\
&\quad \left.\left. + R^D \frac{R^Y - p^1_{\theta,y}(D, Z, X)}{p^1_{\theta,y}(D, Z, X)(1 - p^1_{\theta,y}(D, Z, X))} \dot{p}^1_{\theta,y}(Z, X)\right)\right\} \\
&= E\left\{Z\left(Y - g(D, X; \beta)\right) \times \left(s_\theta(Z, X) + s_\theta(Y, D|Z, X) + \frac{R^D - p_{\theta,d}(Z, X)}{p_{\theta,d}(Z, X)(1 - p_{\theta,d}(Z, X))} \dot{p}_{\theta,d}(Z, X)\right.\right. \\
&\quad \left.\left. + R^D \frac{R^Y - p^1_{\theta,y}(D, Z, X)}{p^1_{\theta,y}(D, Z, X)(1 - p^1_{\theta,y}(D, Z, X))} \dot{p}^1_{\theta,y}(D, Z, X)\right)\right\} \\
&= E\left[m_{full}(Z, X, D, Y; \beta)\left(s(Z, X)' + s(D|Z, X)' + s(Y|D, Z, X)'\right)\right]
\end{aligned}
$$

The second equality follows law of iterated expectation and that $\frac{E\left[R^D R^Y | Z, X, D, Y\right]}{p_{11}} = 1$; the last equality follows the definition of m_{full} and the fact the SMAR assumption such that

$$
\begin{aligned}
E\left[R^D | Z, X, D, Y\right] - p_{\theta,d}(Z, X) &= 0 \\
E\left[R^Y | Z, X, D, Y, R^D = 1\right] - p^1_{\theta,y}(D, Z, X) &= 0
\end{aligned}
$$

Then we need to show, the expectation of interaction between $\mathcal{S}_O$ and the augmenting terms in φ equals to zero. For the second term, we have:

$$
\begin{aligned}
&E\left\{\left(\frac{R^Y}{p_y}-\frac{R^DR^Y}{p_{11}}\right)Z\left[(Y-E\left[g(D,X;\beta)|Z,X\right])-(E\left[Y|Z,X\right]-E\left[g(D,X;\theta)|Z,X\right])\right]\mathcal{S}_O\right\}\\
&=E\left\{(1-p_d)\left(\frac{(1-R^D)R^Y}{p_{01}}-\frac{R^DR^Y}{p_{11}}\right)\right.\\
&\quad\times Z\left[(Y-E\left[g(D,X;\beta)|Z,X\right])-(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right])\right]\\
&\quad\left.\times\left(R^DR^Ys_\theta(Y,D|Z,X)+(1-R^D)R^Ys_\theta(Y|Z,X)\right)\right\}\\
&=E\left[(1-p_d)Z\left[(Y-E\left[g(D,X;\beta)|Z,X\right])-(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right])\right]\right.\\
&\quad\left.\times(s_\theta(Y|Z,X)-s_\theta(Y,D|Z,X))\right]\\
&=E\left[(1-p_d)Z\left[(Y-E\left[g(D,X;\beta)|Z,X\right])-(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right])\right]s_\theta(D|Z,X,Y)\right]\\
&=0
\end{aligned}
$$

The third term equals to

$$
\begin{aligned}
&E\left\{\left(\frac{R^D}{p_d}-\frac{R^DR^Y}{p_{11}}\right)Z\left[(E\left[Y|D,Z,X\right]-g(D,X;\beta))-(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right])\right]\mathcal{S}_O\right\}\\
&=E\left\{\left(\frac{R^D}{p_{11}+p_{10}}-\frac{R^DR^Y}{p_{11}}\right)\right.\\
&\quad\times Z\left[(E\left[Y|D,Z,X\right]-g(D,X;\beta))-(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right])\right]\\
&\quad\left.\times\left(R^DR^Ys_\theta(Y,D|Z,X)+R^D(1-R^Y)s_\theta(D|Z,X)\right)\right\}\\
&=E\left[\frac{p_{10}}{p_d}\tilde{Z}\left[(E\left[Y|D,Z,X\right]-g(D,X;\beta))-(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right])\right]\right.\\
&\quad\left.\times(s_\theta(D|Z,X)-s_\theta(Y,D|Z,X))\right]\\
&=E\left[-\frac{p_{10}}{p_d}Z\left[(E\left[Y|D,Z,X\right]-g(D,X;\beta))-(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right])\right]\right.\\
&\quad\left.\times s_\theta(Y|D,Z,X)=0\right.
\end{aligned}
$$

The last term equals to

$$
\begin{aligned}
& E\left\{\left(1-\frac{R^D R^Y}{p_{11}}\right) Z\left(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right]\right)\mathcal{S}_O\right. \\
& \left.\quad \times Z\left(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right]\right)\right\} \\
& = E\left\{(-p_{01}s_\theta(D|Z,X,Y)-p_{10}s_\theta(Y|Z,X,D)-s_\theta(Y,D|Z,X))\right. \\
& \quad \left.\times Z\left(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right]\right)\right\} \\
& = 0
\end{aligned}
$$

Then, we confirm that φ is in the tangent set, we can rewrite φ into the form:

$$
\begin{aligned}
\varphi = {} & R^D R^Y\left\{\frac{1}{p_{11}}\left[Z(Y-g(D,X;\beta))-Z\left(E\left[Y|D,Z,X\right]-g(D,X;\beta)\right)\right]+\right. \\
& \frac{1-p_d}{p_{11}}\left[Z\left(Y-E\left[Y|Z,X\right]\right)\right] \\
& \left.+\frac{1}{p_d}\left[Z\left(E\left[Y|D,Z,X\right]-g(D,X;\beta)\right)-Z\left(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right]\right)\right]\right\} \\
& + R^D(1-R^Y)\left\{\frac{1}{p_d}\left[Z\left(E\left[Y|D,Z,X\right]-g(D,X;\beta)\right)-Z\left(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right]\right)\right]\right\} \\
& + (1-R^D)R^Y\left\{\frac{1-p_d}{p_{01}}Z\left(Y-E\left[Y|Z,X\right]\right)\right\}+Z\left(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right]\right.
\end{aligned}
$$

Set $b_{11}, b_{01}, b_{10}, b_{00}$ to be zero, and we can easily confirm that

$$
\begin{aligned}
& Z\left(E\left[Y|Z,X\right]-E\left[g(D,X;\theta)|Z,X\right]\right)\in L_0^2\left(F(Z,X)\right) \\
& \frac{1-p_d(Z,X)}{p_{01}(Z,X)}Z\left(Y-E\left[Y|Z,X\right]\right)\in L_0^2\left(F(Y|Z,X)\right) \\
& \frac{1}{p_d(Z,X)}\left[Z\left(E\left[Y|D,Z,X\right]-g(D,X;\beta)\right)-Z\left(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right]\right)\right] \\
& \in L_0^2\left(F(D|Z,X)\right)
\end{aligned}
$$

The last to confirm is that the function following $R^D R^Y$ is in $L_0^2\left(F(Y, D|Z, X)\right)$, and this is the case when either condition (1) or (2) in Theorem 1.5.6 holds.

Under condition (1):

$$E\left\{\frac{1}{p_{11}}\left[Z(Y-g(D,X;\beta))-Z\left(E\left[Y|D,Z,X\right]-g(D,X;\beta)\right)\right]+\frac{1-p_d}{p_{11}}\left[Z\left(Y-E\left[Y|Z,X\right]\right)\right]\right.$$
$$\left.+\frac{1}{p_d}\left[Z\left(E\left[Y|D,Z,X\right]-g(D,X;\beta)\right)-Z\left(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right]\right)\right]|Z,X\right\}=0$$

because all the missing probabilities depend on the fully observed Z, X.

Under condition (2):

$$\begin{aligned}
&E\left\{\frac{1}{p_{11}}\left[Z(Y-g(D,X;\beta))-Z\left(E\left[Y|D,Z,X\right]-g(D,X;\beta)\right)\right]+\right.\\
&\left.\frac{1-p_d}{p_{11}}\left[Z\left(Y-E\left[Y|Z,X\right]\right)\right]|Z,X,D\right\}\\
&=E\left\{\frac{1}{p_{11}}\left[Z(Y-g(D,X;\beta))-Z\left(E\left[Y|D,Z,X\right]-g(D,X;\beta)\right)\right]\right.\\
&\left.+\frac{1-p_d}{p_{11}}\left[Z\left(Y-E\left[Y|D,Z,X\right]\right)\right]|Z,X,D\right\}\\
&=0
\end{aligned}$$

and can be interpreted as a function in $L_0^2\left(F\left(Y|D,Z,X\right)\right)$.

$$E\left[\frac{1}{p_d}\left[Z\left(E\left[Y|D,Z,X\right]-g(D,X;\beta)\right)-Z\left(E\left[Y|Z,X\right]-E\left[g(D,X;\beta)|Z,X\right]\right)\right]|Z,X\right]=0$$

and can be interpreted as a function in $L_0^2\left(F\left(D|Z,X\right)\right)$.

Note that $s_\theta(Y, D|Z, X) = s_\theta(Y|D, Z, X) + s_\theta(D|Z, X)$. Therefore, $\varphi \in \mathcal{T}$.

Therefore, given A, the efficient influence function is $-(AG)^{-1}A\varphi$, and the variance of it is $(AG)^{-1}AV_{MAR}A'(AG)^{-1'}$. The efficient influence function

involves A affecting the variance, and we choose the variance minimizer to be $A = G'V_{MAR}^{-1}$, then the efficiency bound is $\Omega = (G'V_{MAR}^{-1}G)^{-1}$.

Appendix B

Appendix for Chapter 2

B.1 Examples of the Target Parameters

Table B.1 contains the list of target parameters. The table is taken from Mogstad, Santos and Torgovitsky (2018).

B.2 More Discussions

B.2.1 Point-wise and Uniform Sharp Bounds on MTE

In Section 2.2, we provided some examples of target parameters. The building block for these parameters is the MTE, $m_1(u) - m_0(u)$ (suppressing x). Heckman and Vytlacil (2005) show why this fundamental parameter can be of independent interest. Unlike other target parameters proposed here, we may want to allow the MTE to be a function of u (beyond evaluating it at a fixed u). In this section, we discuss the subtle issue of point-wise and uniform sharp bounds on $\tau_{MTE}(u) \equiv m_1(u) - m_0(u)$ as a function of u.

Suppress X for simplicity. Recall $q(u) \equiv \{q(e|u)\}_{e\in\mathcal{E}}$ and $\mathcal{Q} \equiv \{q(\cdot) : \sum_e q(e|u) = 1\,\forall u \text{ and } q(e|u) \geq 0\,\forall(e,u)\}$. Let $\mathcal{M}$ be the set of MTE functions, i.e.,

$$\mathcal{M} \equiv \Big\{m_1(\cdot) - m_0(\cdot) : m_d(\cdot) = E[Y_d|U=\cdot] = \sum_{e\in\mathcal{E}:g_e(d)=1} q(e|\cdot)\,\forall d \in \{0,1\} \text{ for } q(\cdot) \in \mathcal{Q}\Big\}.$$

Target Parameters	Expressions	Ranges of u	Weights
			$w_d(u,z,x)$
Average Treatment Effect (ATE)	$E[Y(1)-Y(0)]$	$[0,1]$	$2d-1$
LATE for Compliers (LATE-C) given $x \in \mathcal{X}$	$E\{Y(1)-Y(0)\|u \in [P(z_0,x),P(z_1,x)]\}$	$[P(z_0,x),P(z_1,x)]$	$(2d-1)\times\frac{1(u\in[P(z_0,x),P(z_1,x)])}{P(z_1,x)-P(z_0,x)}$
LATE for Always-Takers (LATE-AT) given $x \in \mathcal{X}$	$E\{Y(1)-Y(0)\|u \in [0,P(z_0,x)]\}$	$[0,P(z_0,x)]$	$(2d-1)\times\frac{1(u\in[0,P(z_0,x)])}{P(z_0,x)}$
LATE for Never Takers (LATE-NT) given $x \in \mathcal{X}$	$E\{Y(1)-Y(0)\|u \in [P(z_1,x),1]\}$	$[P(z_1,x),1]$	$(2d-1)\times\frac{1(u\in[P(z_1,x),1])}{1-P(z_1,x)}$
LATE for $[\underline{u},\overline{u}]$	$E[Y(1)-Y(0)\|u \in [\underline{u},\overline{u}]]$	$[P(z_0,x),P(z_1,x)]$	$(2d-1)\times\frac{1(u\in[\underline{u},\overline{u}])}{\overline{u}-\underline{u}}$
Marginal Treatment Effect (MTE)*	$E[Y(1)-Y(0)\|u']$	u'	$(2d-1)\times 1(u=u')$
Policy Relevant Treatment Effect (PRTE) for a new policy (P',Z')	$\frac{E(Y')-E(Y)}{E(D')-E(D)}$	$[0,1]$	$(2d-1)\times\frac{\Pr[u\le P'(z')]-\Pr[u\le P'(z)]}{E[P(Z')]-E[P(Z)]}$

* The MTE uses the Dirac measure at u', while the other target parameters use the Lebesgue measure on $[0,1]$.

Table B.1Examples of the Target Parameters

The bounds on $\tau_{MTE} \in \mathcal{M}$ in the ∞-LP are given by using a Dirac delta function as a weight. Therefore, given evaluation point $u \in [0,1]$, (∞-LP1)–(∞-LP3) can be simplified as follows, defining the upper and lower bounds $\overline{\tau}(u)$ and $\underline{\tau}(u)$ (being explicit about the evaluation point) on $\tau_{MTE}(u)$:

$$\overline{\tau}(u) = \sup_{q\in\mathcal{Q}} \sum_{e\in\mathcal{E}:g_e(1)=1} q(e|u) - \sum_{e\in\mathcal{E}:g_e(0)=1} q(e|u) \tag{B.2.1}$$

$$\underline{\tau}(u) = \inf_{q\in\mathcal{Q}} \sum_{e\in\mathcal{E}:g_e(1)=1} q(e|u) - \sum_{e\in\mathcal{E}:g_e(0)=1} q(e|u) \tag{B.2.2}$$

subject to

$$\sum_{e:g_e(d)=1} \int_{\mathcal{U}_z^d} q(e|\tilde{u})d\tilde{u} = p(1,d|z) \qquad \forall (d,z) \in \{0,1\}^2. \tag{B.2.3}$$

Then, for any fixed $u \in [0,1]$,

$$\underline{\tau}(u) \le \tau_{MTE}(u) \le \overline{\tau}(u).$$

We argue that these bounds are point-wise sharp but not necessarily uniformly sharp for $\tau_{MTE}(\cdot)$.[1]

Definition B.2.1 (Point-wise Sharpness)**.** *$\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ are point-wise sharp if, for any $\bar{u} \in [0,1]$, there exist $\overline{\tau}_{MTE,\bar{u}}, \underline{\tau}_{MTE,\bar{u}} \in \mathcal{M}$ such that $\overline{\tau}(\bar{u}) = \overline{\tau}_{MTE,\bar{u}}(\bar{u})$ and $\underline{\tau}(\bar{u}) = \underline{\tau}_{MTE,\bar{u}}(\bar{u})$.*

Theorem B.2.1. *$\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ are point-wise sharp bounds on $\tau_{MTE}(\cdot)$.*

The proofs of this and other theorems appear later. Note that point-wise bounds will maintain some properties of an MTE function, but not all. For uniform sharpness, $\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ themselves have to be MTE functions on $[0,1]$, i.e., $\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ should be elements in $\mathcal{M}$.

Definition B.2.2 (Uniform Sharpness)**.** *$\overline{\tau}(\cdot)$ and $\underline{\tau}(\cdot)$ are uniformly sharp if $\overline{\tau}(\cdot), \underline{\tau}(\cdot) \in \mathcal{M}$.*

The following theorem is almost immediate.

Theorem B.2.2. *$\overline{\tau}(\cdot)$ is uniformly sharp if and only if there exists $q^*(\cdot) \in \mathcal{Q}$ such that $q^*(\cdot)$ is in the feasible set and $\overline{\tau}(u) = \sum_{e \in \mathcal{E}: g_e(1)=1} q^*(e|u) - \sum_{e \in \mathcal{E}: g_e(0)=1} q^*(e|u)$ for all $u \in [0,1]$. Similarly, $\underline{\tau}(\cdot)$ is uniformly sharp if and only if there exists $q^\dagger(\cdot) \in \mathcal{Q}$ such that $q^\dagger(\cdot)$ is in the feasible set and $\underline{\tau}(u) = \sum_{e \in \mathcal{E}: g_e(1)=1} q^\dagger(e|u) - \sum_{e \in \mathcal{E}: g_e(0)=1} q^\dagger(e|u)$ for all $u \in [0,1]$.*

[1] See Sergio Firpo and Geert Ridder (2019) for related definitions of point-wise and uniform sharpness.

The following is a more useful result that relates point-wise bounds with uniform bounds. For each $\bar{u}$, let $q^*_{\bar{u}}(\cdot)$ and $q^\dagger_{\bar{u}}(\cdot)$ be the point-wise maximizer and minimizer of (B.2.1)–(B.2.3), respectively.

Corollary B.2.1. *$\overline{\tau}(\cdot)$ is uniformly sharp if and only if there exists $q^*(\cdot) \in \mathcal{Q}$ such that $q^*(\cdot)$ is in the feasible set and $q^*_{\bar{u}}(\bar{u}) = q^*(\bar{u})$ for all $\bar{u} \in [0,1]$. Also, $\underline{\tau}(u)$ is uniformly sharp if and only if there exists $q^\dagger(\cdot) \in \mathcal{Q}$ such that $q^\dagger(\cdot)$ is in the feasible set and $q^\dagger_{\bar{u}}(\bar{u}) = q^\dagger(\bar{u})$ for all $\bar{u} \in [0,1]$.*

Based on the Bernstein approximation we introduce, this corollary implies that for a uniform upper bound to exist, there should exist a common maximizer θ^* such that θ^* is in the feasible set of the LP and $\overline{\tau}(u) = \sum_{k\in\mathcal{K}} \left\{ \sum_{e\in\mathcal{E}:g_e(1)=1} \theta^{e*}_k b_k(u) - \sum_{e\in\mathcal{E}:g_e(0)=1} \theta^{e*}_k b_k(u) \right\}$ *for all u*. In other words, if $\theta^*_{\bar{u}}$ is the maximizer of the LP for given $\bar{u}$, then there should exist θ^* in the feasible set such that $\theta^*_{\bar{u}} = \theta^*$ for all $\bar{u} \in [0,1]$. Since this condition will not generally hold, uniformly sharp bounds on the MTE may not exist. The condition can be verified in practice by implementing the LP in a finite grid of u in $[0,1]$ and checking whether θ^*_u is constant for all values in the grid.

B.2.2 Inference

It is important to construct a confidence set for our target parameter or its bounds in order to account for the sampling variation in measuring treatment effectiveness. It will also be interesting to develop a procedure to conduct a specification test for the identifying assumptions discussed in Section 2.6. The problem of statistical inference when the identified set is constructed

via linear programming has been studied in, e.g.,Deb et al. (2017), Mogstad, Santos and Torgovitsky (2018), Yu-Wei Hsieh, Xiaoxia Shi and Matthew Shum (2018), and Alexander Torgovitsky (2019*b*) . Among these papers, Magne Mogstad, Andres Santos and Alexander Torgovitsky (2017)'s setting is closest to ours, and their inference procedure can be directly adapted to our problem. Instead of repeating their result here, we only briefly discuss the procedure.

Recall $q(u,x) \equiv \{q(e|u,x)\}_{e\in\mathcal{E}}$ is the latent distribution and $p \equiv \{p(1,d|z,x)\}_{d,z,x}$ is the distribution of the data, and R_τ, R_0, R_1, and R_2 denote the linear operators of $q(\cdot)$ that correspond to the target and constraints. Consider the following hypotheses:

$$H_0 : p \in \mathcal{P}_0, \qquad H_1 : p \in \mathcal{P}\backslash\mathcal{P}_0,$$

where

$$\mathcal{P}_0 \equiv \{p \in \mathcal{P} : Rq = a \text{ for some } q \in \mathcal{Q}\}$$

and

$$R \equiv (R_\tau', R_0', R_1', R_2')'$$
$$a \equiv (\tau, p', a_1', a_2')'$$

Suppose $\hat{R}$ and $\hat{a}$ are sample counterparts of R and a. Then, a minimum distance test statistic can be constructed as

$$T_n(\tau) \equiv \inf_{q\in\mathcal{Q}_K} \sqrt{n} \left\| \hat{R}q - \hat{a} \right\|.$$

Similar to Mogstad, Santos and Torgovitsky (2017), $T_n(\tau)$ is the solution to a convex optimization problem that can be reformulated as an LP using duality. A $(1-\alpha)$-confidence set for the target parameter τ can be constructed by inverting the test:

$$CS_{1-\alpha} \equiv \{\tau : T_n(\tau) \leq \hat{c}_{1-\alpha}\}$$

where $\hat{c}_{1-\alpha}$ is the critical value for the test. The resulting object is of independent interest, and it can further be used to conduct specification tests. The large sample theory for $T_n(\tau)$, as well as a bootstrap procedure to calculate $\hat{c}_{1-\alpha}$, will directly follow according to Mogstad, Santos and Torgovitsky (2017), which is omitted for succinctness.

B.2.3 Linear Programming with Continuous X

Suppose X is continuously distributed and assume $\mathcal{X} = [0,1]^{d_X}$. Let $q(u,x) \equiv \{q(e|u,x)\}_{e\in\mathcal{E}}$ and $p(x) \equiv \{p(1,d|z,x)\}_{d,z}$. Recall that $R_\tau : \mathcal{Q} \to \mathbb{R}$ and $R : \mathcal{Q} \to \mathbb{R}^{d_p}$ are the linear operators of $q(\cdot)$ where d_p is the dimension of p. Consider the following LP:

$$\overline{\tau} = \sup_{q\in\mathcal{Q}} R_\tau q, \tag{B.2.4}$$

$$\underline{\tau} = \inf_{q\in\mathcal{Q}} R_\tau q, \tag{B.2.5}$$

$$s.t. \quad (Rq(x) = p(x) \qquad \text{for all } x \in \mathcal{X}, \tag{B.2.6}$$

where $(Rq)(x) = p(x)$ emphasizes the dependence on x, and thus contains infinitely many constraints. Therefore, this LP is infinite dimensional because

of not only the decision variable but also the constraints. The problem with q is addressed with the sieve approximation. To address the problem with continuous X, we proceed as follow. Note that, in general, $E\left|h(X)\right| = 0$ if and only if $h(x) = 0$ almost everywhere in $\mathcal{X}$. Therefore, each j-th equation in the equality restrictions (B.2.6) can be replaced by

$$E\left|(Rq)_j(X) - p_j(X)\right| = 0.$$

Now, for the sieve space of $\mathcal{Q}$, we consider

$$\tilde{\mathcal{Q}}_K \equiv \left\{ \left\{ \sum_{k=1}^{\tilde{K}} \theta_k^e b_k(u,x) \right\}_{e\in\mathcal{E}} : \sum_{e\in\mathcal{E}} \theta_k^e = 1 \,\forall k \in \tilde{\mathcal{K}} \text{ and } \theta_k^e \geq 0 \,\forall (e,k) \right\} \subseteq \mathcal{Q}, \tag{B.2.7}$$

where $b_k(u,x)$ is a bivariate Bernstein polynomial and $\tilde{\mathcal{K}} \equiv \{1, ..., \tilde{K}\}$. Then,

$$\begin{aligned} E[\tau_d(Z,X)] &= \sum_{e:g_e(d)=1} \sum_{k\in\tilde{\mathcal{K}}} \theta_k^e \int E[b_k(u,X) w_d(u,Z,X)] du \\ &\equiv \sum_{e:g_e(d)=1} \sum_{k\in\tilde{\mathcal{K}}} \theta_k^e \tilde{\gamma}_k^d, \end{aligned} \tag{B.2.8}$$

where $\tilde{\gamma}_k^d \equiv \int E[b_k(u,X) w_d(u,Z,X)] du$. Also,

$$\begin{aligned} p(y,d|z,x) &= \sum_{e:g_e(d)=y} \sum_{k\in\tilde{\mathcal{K}}} \theta_k^e \int_{\mathcal{U}_{z,x}^d} b_k(u,x) du \\ &\equiv \sum_{e:g_e(d)=y} \sum_{k\in\tilde{\mathcal{K}}} \theta_k^e \tilde{\delta}_k^d(z,x), \end{aligned} \tag{B.2.9}$$

where $\tilde{\delta}_k^d(z,x) \equiv \int_{\mathcal{U}_{z,x}^d} b_k(u,x) du$. Let $\tilde{\theta} \equiv \{\theta_k^e\}_{(e,k)\in\mathcal{E}\times\tilde{\mathcal{K}}}$ and let

$$\tilde{\Theta}_{\tilde{K}} \equiv \left\{ \tilde{\theta} : \sum_{e\in\mathcal{E}} \theta_k^e = 1 \,\forall k \in \tilde{\mathcal{K}} \text{ and } \theta_k^e \geq 0 \,\forall (e,k) \right\}.$$

Then, we can formulate the following finite-dimensional LP:

$$\overline{\tau}_{\tilde{K}} = \max_{\theta \in \Theta_{\tilde{K}}} \sum_{k \in \tilde{\mathcal{K}}} \Big\{ \sum_{e: g_e(1)=1} \theta_k^e \tilde{\gamma}_k^1 - \sum_{e: g_e(0)=1} \theta_k^e \tilde{\gamma}_k^0 \Big\} \tag{B.2.10}$$

$$\underline{\tau}_{\tilde{K}} = \min_{\theta \in \Theta_{\tilde{K}}} \sum_{k \in \tilde{\mathcal{K}}} \Big\{ \sum_{e: g_e(1)=1} \theta_k^e \tilde{\gamma}_k^1 - \sum_{e: g_e(0)=1} \theta_k^e \tilde{\gamma}_k^0 \Big\} \tag{B.2.11}$$

subject to

$$E \left| \sum_{e: g_e(d)=1} \sum_{k \in \tilde{\mathcal{K}}} \theta_k^e \tilde{\delta}_k^d(Z, X) - p(1, d|Z, X) \right| = 0. \tag{B.2.12}$$

In estimation, we use the sample counterparts $\hat{\tilde{\gamma}}_k^d$ and $\hat{\tilde{\delta}}_k^d$ for $\tilde{\gamma}_k^d$ and $\tilde{\delta}_k^d$, and (B.2.12) can be estimated with slackness by

$$\frac{1}{n} \sum_{i=1}^{n} \left| \sum_{e: g_e(d)=1} \sum_{k \in \tilde{\mathcal{K}}} \theta_k^e \hat{\tilde{\delta}}_k^d(Z_i, X_i) - \hat{p}(1, d|Z_i, X_i) \right| \leq \eta,$$

where $\hat{p}(1, d|z, x)$ is some preliminary estimate of $p(1, d|z, x)$ and η is the slackness parameter.

Later, we want to introduce additional constraints from some identifying assumptions:

$$R_1 q = a_1 \tag{B.2.13}$$

$$R_2 q \leq a_2 \tag{B.2.14}$$

For the equality restrictions, we can use the same approach that transforms (B.2.6). For the inequality restrictions (B.2.14), we can allow any identifying assumptions for which R_2 is a matrix rather than an operator:

Assumption MAT. *R_2 is a $\dim(a_2) \times \dim(q)$ matrix.*

Assumptions M and C and the unconditional version of Assumption MTS satisfy this condition.

B.2.4 Equivalence with the IV-Like Estimands

We draw a connection between our approach and the approach used in Mogstad, Santos and Torgovitsky (2018). In particular, we show that the identified set of the MTR functions $\mathcal{M}_{id}$ used in Mogstad, Santos and Torgovitsky (2018) is equivalent to the set of MTR functions derived from the feasible set used in this paper. Therefore, the feasible set in this paper contains no less information about the data than those contained in $\mathcal{M}_{id}$ via IV-like estimands in their paper.

The IV-like estimand is defined in Proposition 3 in Mogstad, Santos and Torgovitsky (2018), and is stated as below.

Proposition B.2.1 (IV-like Estimand from Mogstad, Santos and Torgovitsky (2018))**.** *Suppose that $s : \{0,1\} \times \mathbf{R}^{d_z \times d_x} \to \mathbf{R}$ is an identified (or known) function that is measurable and has a finite second moment. We refer to such a function s as an IV-like specification and to $\beta_s \equiv E\Big[s(D,Z,X)Y\Big]$ as an IV-like estimand. If (Y,D) are generated according to Assumption SEL and Assumption EX, then*

$$\beta_s = E\Big[\int_0^1 m_0(u,X)\omega_{0s}(u,Z,X)du\Big] + E\Big[\int_0^1 m_1(u,X)\omega_{1s}(u,Z,X)du\Big], \tag{B.2.15}$$

where $\omega_{0s}(u,z,x) = s(0,z,x)1[u > p(z,x)]$, *and* $\omega_{1s}(u,z,x) = s(1,z,x)1[u \leq p(z,x)]$.

For the MTR functions to be consistent with the data, the following conditions need to be satisfied:

$$E[Y|D=0,Z,X] = E[Y_0|U > p(Z,X),Z,X] = \frac{1}{1-P(Z,X)}\int_{p(Z,X)}^{1} m_0(u,X)du, \tag{B.2.16}$$

$$E[Y|D=1,Z,X] = E[Y_1|U \leq p(Z,X),Z,X] = \frac{1}{P(Z,X)}\int_{0}^{p(Z,X)} m_1(u,X)du. \tag{B.2.17}$$

Define the identified set as:

$$\mathcal{M}_{id} = \Big\{m = (m_0,m_1), m_0, m_1 \in L^2 : m_0, m_1 \text{ satisfies equation (B.2.16) and (B.2.17) a.s}\Big\}.$$

This identified set is defined in Mogstad, Santos and Torgovitsky (2018, Section 2.5). The definition follows the fact that the MTR functions in $\mathcal{M}_{id}$ are compatible with the observed conditional means of Y. In this sense, it exhausts the information of the data contained in the conditional means. When Y is binary, the conditional means of Y contain the information of the complete distribution.

Define the feasible set $\mathcal{Q}_f$ as

$$\mathcal{Q}_f = \Big\{q \in L^2 : q \in \mathcal{Q} \text{ and satisfies equation } (\infty\text{-LP3})\Big\}.$$

To establish the connection with $\mathcal{M}_{id}$, we construct the set of MTR functions based on the feasible set:

$$\mathcal{M}_f = \Big\{ m = (m_0, m_1) : m_d = \sum_{e:g_e(d)=1} q(e|u,x), d = \{0,1\}, q \in \mathcal{Q}_f \Big\}.$$

Then the following holds, proof of which appears later:

Theorem B.2.3. *Suppose Y is discretely distributed. Under the Assumption SEL and EX, $\mathcal{M}_f = \mathcal{M}_{id}$.*

Proposition 3 in Mogstad, Santos and Torgovitsky (2018) shows an equivalence relationship between the identified set $\mathcal{M}_{id}$ and the set of MTR functions satisfying constraints based on selected IV-like estimands. Theorem B.2.3 shows that the information contained in our feasible set used in the LP is the same as the selected IV-like estimands that exhaust the available information. Theorem B.2.3 can be extended to the case where Y is discrete and X is continuous. When Y is a non-binary discrete outcome variable, $\mathcal{M}_{id}$ and $\mathcal{M}_f$ only exhaust the information on the conditional means, but not other distributional information. Nonetheless, that missing information is captured by $\mathcal{Q}_f$ that we use as our constraint set, because $q(e|u)$ is defined as the conditional probability of Y taking each value.

B.3 Proofs

B.3.1 Proof of Lemma 2

Fix (d, z, x). By $\sum_{e \in \mathcal{E}} q(e|u,x) = 1$ for $q \in \mathcal{Q}$, we have

$$1 = \sum_{e \in \mathcal{E}} q(e|u,x) = \sum_{e:g_e(d)=1} q(e|u,x) + \sum_{e:g_e(d)=0} q(e|u,x).$$

Then, in (∞-LP3), the constraint with $p(0,d|z,x)$ can be written as

$$p(0,d|z,x) = \int_{\mathcal{U}^d_{z,x}} \sum_{e:g_e(d)=0} q(e|u,x)du = \int_{\mathcal{U}^d_{z,x}} \Big\{1 - \sum_{e:g_e(d)=1} q(e|u,x)\Big\}du$$
$$= \Pr[D=d|Z=z,X=x] - \int_{\mathcal{U}^d_{z,x}} \sum_{e:g_e(d)=1} q(e|u,x)du.$$

Then by rearranging terms, this constraint becomes

$$p(1,d|z,x) = \int_{\mathcal{U}^d_{z,x}} \sum_{e:g_e(d)=1} q(e|u,x)du,$$

since $\Pr[D=d|Z=z,X=x] - p(0,d|z,x) = p(1,d|z,x)$. Therefore, the constraint with $p(0,d|z,x)$ does not contribute to the restrictions imposed by (∞-LP3) and $q \in \mathcal{Q}$. $\square$

B.3.2 Proof of Theorem 2.5.1

In proving the claim of the theorem, note that Z can be fixed at a certain value, so we fix $Z=z$ here. We first prove with Case (a). To simplify notation, let $q(e_1,...,e_J|u) \equiv \Pr[\epsilon \in \{e_1,...,e_J\}|u] = \sum_{j=1}^J q(e_j|u)$. Based on Table (2.1), we can easily derive

$$p(1,1|z,1) = \int_0^{P(z)} \sum_{e:g_e(1,1)=1} q(e|u)du = \int_0^{P(z)} q(9,...,16|u)du,$$
$$p(1,1|z,0) = \int_0^{P(z)} \sum_{e:g_e(1,0)=1} q(e|u)du = \int_0^{P(z)} q(5,...,8,13,...,16|u)du,$$
$$p(1,0|z,1) = \int_{P(z)}^1 \sum_{e:g_e(0,1)=1} q(e|u)du = \int_{P(z)}^1 q(3,4,7,8,11,12,15,16|u)du,$$
$$p(1,0|z,0) = \int_{P(z)}^1 \sum_{e:g_e(0,0)=1} q(e|u)du = \int_{P(z)}^1 q(2,4,6,8,10,12,14,16|u)du.$$

Define the operator

$$T_z^d q^e \equiv \int_{\mathcal{U}_z^d} q(e|u)du.$$

Then, for the r.h.s. $(p_{11|z1}, p_{11|z0}, p_{10|z1}, p_{10|z0})'$ of the constraints in (LP_W3) that correspond to $Z = z$, the corresponding l.h.s. is

$$\begin{pmatrix} \int_0^{P(z)} q(9,...,16|u)du \\ \int_0^{P(z)} q(5,...,8,13,...,16|u)du \\ \int_{P(z)}^1 q(3,4,7,8,11,12,15,16|u)du \\ \int_{P(z)}^1 q(2,4,6,8,10,12,14,16|u)du \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & T_z^1 \\ 0 & 0 & 0 & 0 & T_z^1 & T_z^1 & T_z^1 & T_z^1 & 0 & 0 & 0 & 0 & T_z^1 & T_z^1 & T_z^1 & T_z^1 \\ 0 & 0 & T_z^0 & T_z^0 & 0 & 0 & T_z^0 & T_z^0 & 0 & 0 & T_z^0 & T_z^0 & 0 & 0 & T_z^0 & T_z^0 \\ 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 & 0 & T_z^0 \end{pmatrix} q$$

$$\equiv Tq,$$

where T is a matrix of operators implicitly defined and $q(u) \equiv (q(1|u), ..., q(16|u))$. Now for $q \in \mathcal{Q}_K$, define a $16K$-vector

$$\theta \equiv \begin{pmatrix} \theta^1 \\ \vdots \\ \theta^{16} \end{pmatrix}$$

where, for each $e \in \{1, ..., 16\}$, $\theta^e \equiv (\theta_1^e, ..., \theta_K^e)'$. Similarly, let $b(u) \equiv (b_1(u), ..., b_K(u))'$. Then, we have $q(e|u) = b(u)'\theta^e$. Let H be a 16×16 diagonal matrix of 1's and 0's that imposes additional identifying assumptions on the outcome data-generating process. In this proof, H is used to incorporate Assumption R(i). Given H, the constraints in (LP_W3) (that correspond to $Z = z$) can be written as

$$THq = \{TH \otimes b'\}\,\theta = (p_{11|z1}, p_{11|z0}, p_{10|z1}, p_{10|z0})'.$$

Now, we prove the claim of the theorem. Suppose the claim is not true, i.e., the even rows are linearly dependent to odd rows in TH. Given the form of T, which has full rank under Assumption R(ii)(a), this linear dependence only occurs when H is such that $H_{jj} = 1$ for $j \in \{1, 4, 13, 16\}$ and 0 otherwise. But, according to Table 2.1, this implies that $\Pr[Y(d, w) \neq Y(d, w')] = 0$ for all d and $w \neq w'$, which contradicts Assumption R(i). This proves the theorem for Case (a).

Now we move to prove the theorem for Case (b), analogous to the previous case. For every z, we can derive

$$p(1,1|z,1) = \int_0^{P(z,1)} \sum_{e:g_e(1,1)=1} q(e|u)du = \int_0^{P(z,1)} q(9,...,16|u)du,$$
$$p(1,1|z,0) = \int_0^{P(z,0)} \sum_{e:g_e(1,0)=1} q(e|u)du = \int_0^{P(z,0)} q(5,...,8,13,...,16|u)du,$$
$$p(1,0|z,1) = \int_{P(z,1)}^1 \sum_{e:g_e(0,1)=1} q(e|u)du = \int_{P(z,1)}^1 q(3,4,7,8,11,12,15,16|u)du,$$
$$p(1,0|z,0) = \int_{P(z,0)}^1 \sum_{e:g_e(0,0)=1} q(e|u)du = \int_{P(z,0)}^1 q(2,4,6,8,10,12,14,16|u)du.$$

Define

$$T^d_{z,w} q^e \equiv \int_{\mathcal{U}^d_{z,w}} q(e|u)du$$

where $\mathcal{U}^d_{z,w}$ can be analogously defined. Then,

$$\begin{pmatrix} \int_0^{P(z,w)} q(9,...,16|u)du \\ \int_0^{P(z,w')} q(5,...,8,13,...,16|u)du \\ \int_{P(z,w)}^1 q(3,4,7,8,11,12,15,16|u)du \\ \int_{P(z,w')}^1 q(2,4,6,8,10,12,14,16|u)du \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} & T^1_{z,w} \\ 0 & 0 & 0 & 0 & T^1_{z,w'} & T^1_{z,w'} & T^1_{z,w'} & T^1_{z,w'} & 0 & 0 & 0 & 0 & T^1_{z,w'} & T^1_{z,w'} & T^1_{z,w'} & T^1_{z,w'} \\ 0 & 0 & T^0_{z,w} & T^0_{z,w} & 0 & 0 & T^0_{z,w} & T^0_{z,w} & 0 & 0 & T^0_{z,w} & T^0_{z,w} & 0 & 0 & T^0_{z,w} & T^0_{z,w} \\ 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} & 0 & T^0_{z,w'} \end{pmatrix} q$$

$$\equiv \tilde{T}q,$$

where $\tilde{T}$ is a matrix of operators implicitly defined. Then, inserting H, the constraint becomes

$$\tilde{T}Hq = \left\{\tilde{T}H \otimes b'\right\}\theta = (p_{11|z1}, p_{11|z0}, p_{10|z1}, p_{10|z0})'.$$

Then the remaining argument is the same as in the previous case, which completes the proof. □

B.3.3 Proof of Theorem B.2.1

For any given $\bar{u} \in [0,1]$, $\overline{\tau}(\bar{u}) = \sum_{e\in\mathcal{E}:g_e(1)=1} q^*_{\bar{u}}(e|\bar{u}) - \sum_{e\in\mathcal{E}:g_e(0)=1} q^*_{\bar{u}}(e|\bar{u})$ for some $q^*_{\bar{u}}(\cdot) \equiv \{q^*_{\bar{u}}(e|\cdot)\}_{e\in\mathcal{E}}$ in the feasible set of the LP, (B.2.1) and (B.2.3). Therefore, $\overline{\tau}(\bar{u}) = \overline{\tau}_{MTE,\bar{u}}(\bar{u})$ for $\overline{\tau}_{MTE,\bar{u}}(\bar{u}) = \sum_{e\in\mathcal{E}:g_e(1)=1} q^*_{\bar{u}}(e|\bar{u}) - \sum_{e\in\mathcal{E}:g_e(0)=1} q^*_{\bar{u}}(e|\bar{u})$, which is in $\mathcal{M}$ by definition. We can have a symmetric proof for $\underline{\tau}(\cdot)$. □

B.3.4 Proof of Theorem B.2.2

Again, by the fact that $\tau_{MTE}(\cdot) = \sum_{e\in\mathcal{E}:g_e(1)=1} q(e|\cdot) - \sum_{e\in\mathcal{E}:g_e(0)=1} q(e|\cdot)$ in general, $\overline{\tau}(u) = \sum_{e\in\mathcal{E}:g_e(1)=1} q^*(e|u) - \sum_{e\in\mathcal{E}:g_e(0)=1} q^*(e|u)$ for all $u \in [0,1]$ is equivalent to $\overline{\tau}(\cdot)$ being contained in $\mathcal{M}$, and similarly for $\underline{\tau}(\cdot)$. □

B.3.5 Proof of Theorem B.2.3

From (∞-LP3), we can write $E[Y|D=0,Z,X]$ in terms of $q(e|u,X)$ as below:

$$\begin{aligned} E[Y|D=0,Z,X] &= \Pr[Y=1|D=0,Z,X] = \frac{\Pr[Y=1,D=0|Z,X]}{\Pr[D=0|Z,X]} \\ &= \frac{1}{1-P(Z,X)} \sum_{e:g_e(0)=1} \int_{P(Z,X)}^{1} q(e|u,X)du \\ &= \frac{1}{1-P(Z,X)} \int_{P(Z,X)}^{1} \sum_{e:g_e(0)=1} q(e|u,X)du \end{aligned} \tag{B.3.1}$$

Therefore, for $(m_0, m_1) \in \mathcal{M}_f$

$$E[Y|D=0,Z,X] = \frac{1}{P(Z,X)} \int_{P(Z,X)}^{1} m_0(u,X)du$$

and symmetrically,

$$E[Y|D=1,Z,X] = \frac{1}{P(Z,X)} \int_{0}^{P(Z,X)} m_1(u,X)du$$

We conclude that $\mathcal{M}_f \subset \mathcal{M}_{id}$.

Now suppose $m \in \mathcal{M}_{id}$. By (B.2.16) and (B.3.1), for $\forall z, x$

$$\frac{1}{1-P(z,x)} \int_{P(z,x)}^{1} m_0(u,x)du = \frac{1}{1-P(z,x)} \sum_{e:g_e(0)=1} \int_{P(z,x)}^{1} q(e|u,x)du$$

and,

$$\int_{P(z,x)}^{1} \left[m_0(u,x) - \sum_{e:g_e(0)=1} q(e|u,x) \right] du = 0$$

This equality holds for all the possible values of $P(z,x)$, we conclude that $m_0(u,x) = \sum_{e:g_e(0)=1} q(e|u,x)$ on the support $u \in [0,1]$, $\forall x$ following the fundamental theorem of calculus. Following the symmetric procedure, we can conclude that $m_1(u,x) = \sum_{e:g_e(1)=1} q(e|u,x)$. And we show that $\mathcal{M}_{id} \subset \mathcal{M}_f$. Thus, $\mathcal{M}_f = \mathcal{M}_{id}$.

Appendix C

Appendix for Chapter 3

C.1 Proof of Theorem 2

Now for d=0,1, define

$$p^d_{t-1}(0,0) =: P(\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} = 0 | D = d)$$

$$p^d_{t-1}(0,\neq 0) =: P(\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} \neq 0 | D = d)$$

$$p^d_{t-1}(\neq 0,0) =: P(\overline{Y}^1_{t-1} \neq 0, \overline{Y}^0_{t-1} = 0 | D = d)$$

Without randomization assumption, $E[Y^1_t | \overline{Y}^1_{t-1} = 0]$ is not identified. Now we have

$$ATETS_t = \frac{\overbrace{P(Y_t = 1, \overline{Y}_{t-1} = 0 | D = 1)P(D = 1)}^{A} + \overbrace{P(Y^1_t = 1, \overline{Y}^1_{t-1} = 0 | D = 0)P(D = 0)}^{B}}{P(\overline{Y}_{t-1} = 0 | D = 1)P(D = 1) + P(\overline{Y}^1_{t-1} = 0 | D = 0)P(D = 0)}$$
$$- \frac{\overbrace{P(Y^0_t = 1, \overline{Y}^1_{t-1} = 0 | D = 1)P(D = 1)}^{C} + \overbrace{P(Y^0_t = 1, \overline{Y}^1_{t-1} = 0 | D = 0)P(D = 0)}^{D}}{P(\overline{Y}_{t-1} = 0 | D = 1)P(D = 1) + P(\overline{Y}^1_{t-1} = 0 | D = 0)P(D = 0)} \quad \text{(C.1.1)}$$

In C.1.1, A is identified.

$$\begin{aligned}
B - C - D =& P(Y_t^1 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 = 0, D = 0)p_{t-1}^0(0,0)P(D = 0) \\
&+ P(Y_t^1 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 \neq 0, D = 0)p_{t-1}^0(0,\neq 0)P(D = 0) \\
&- P(Y_t^0 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 = 0, D = 1)p_{t-1}^1(0,0)P(D = 1) \\
&- P(Y_t^0 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 \neq 0, D = 1)p_{t-1}^1(0,\neq 0)P(D = 1) \\
&- P(Y_t^0 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 = 0, D = 0)p_{t-1}^0(0,0)P(D = 0) \\
&- P(Y_t^0 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 \neq 0, D = 0)p_{t-1}^0(0,\neq 0)P(D = 0)
\end{aligned}$$

Under Assumption 2,

$$P(Y_t^0 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 = 0, D = 0) = P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0)$$

$$P(Y_t^0 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 \neq 0, D = 0) = P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)$$

Plug into $B - C - D$, get

$$\begin{aligned}
B - C - D =& P(Y_t^1 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 = 0, D = 0)p_{t-1}^0(0,0)P(D = 0) \\
&+ P(Y_t^1 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 \neq 0, D = 0)p_{t-1}^0(0,\neq 0)P(D = 0) \\
&- P(Y_t^0 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 = 0, D = 1)p_{t-1}^1(0,0)P(D = 1) \\
&- P(Y_t^0 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 \neq 0, D = 1)p_{t-1}^1(0,\neq 0)P(D = 1) \\
&- P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0)p_{t-1}^0(0,0)P(D = 0) \\
&- P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)p_{t-1}^0(0,\neq 0)P(D = 0)
\end{aligned} \tag{C.1.2}$$

Substitute C.1.2 into C.1.1,

$$
\begin{aligned}
ATETS_t = & \frac{P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 1)P(\overline{Y}_{t-1}|D = 1)P(D = 1)}{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1) + P(D = 0)p^0_{t-1}(0, 0) + P(D = 0)p^0_{t-1}(0, \neq 0)} \\
& + \frac{p^0_{t-1}(0, 0)P(D = 0)[P(Y^1_t = 1|\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} = 0, D = 0) - P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0)]}{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1) + P(D = 0)p^0_{t-1}(0, 0) + P(D = 0)p^0_{t-1}(0, \neq 0)} \\
& + \frac{p^0_{t-1}(0, \neq 0)P(D = 0)[P(Y^1_t = 1|\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} \neq 0, D = 0) - P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)]}{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1) + P(D = 0)p^0_{t-1}(0, 0) + P(D = 0)p^0_{t-1}(0, \neq 0)} \\
& - \frac{P(D = 1)p^1_{t-1}(0, 0)P(Y^0_t = 1|\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} = 0, D = 1)}{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1) + P(D = 0)p^0_{t-1}(0, 0) + P(D = 0)p^0_{t-1}(0, \neq 0)} \\
& - \frac{P(D = 1)p^1_{t-1}(0, \neq 0)P(Y^0_t = 1|\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} \neq 0, D = 1)}{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1) + P(D = 0)p^0_{t-1}(0, 0) + P(D = 0)p^0_{t-1}(0, \neq 0)}
\end{aligned}
\tag{C.1.3}
$$

Notice that C.1.3 is decreasing in $P(Y^0_t = 1|\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} = 0, D = 1)$ and $P(Y^0_t = 1|\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} = 0, D \neq 1)$, increasing in $P(Y^0_t = 1|\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} = 0, D = 0)$ and $P(Y^0_t = 1|\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} = 0, D \neq 0)$. Also $p^1_{t-1}(0, \neq 0) + p^1_{t-1}(0, 0)$, thus

$(A15)$

$$
\begin{aligned}
& \leq \frac{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1)P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 1) + P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1)}{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1) + P(D = 0)p^0_{t-1}(0, 0) + P(D = 0)p^0_{t-1}(0, \neq 0)} \\
& - \frac{p^0_{t-1}(0, 0)P(D = 0)P(Y_t = 1|\overline{Y}_{t-1} = 0|D = 0)}{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1) + P(D = 0)p^0_{t-1}(0, 0) + P(D = 0)p^0_{t-1}(0, \neq 0)} \\
& - \frac{p^0_{t-1}(0, \neq 0)P(D = 0)P(Y_t = 1|\overline{Y}_{t-1} \neq 0|D = 0)}{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1) + P(D = 0)p^0_{t-1}(0, 0) + P(D = 0)p^0_{t-1}(0, \neq 0)}
\end{aligned}
$$

Note that the lower bound is decreasing in both $p^0_{t-1}(0,0) \in [0, P(\overline{Y}_{t-1} = 0|D=0)]$ and $p^0_{t-1}(0,\neq 0) \in [0, P(\overline{Y}_{t-1} = 0|D\neq 0)]$, we have:

$$LB_t = -\frac{P(D=1)P(\overline{Y}_{t-1}=0, D=1) + P(Y_t=1|D=0)P(D=0)}{P(D=1)P(\overline{Y}_{t-1}=0|D=1) + P(D=0)} \quad \text{(C.1.4)}$$

Go back to C.1.3, take $P(Y^0_t = 1|\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} = 0, D = 1) = 0$ and $P(Y^0_t = 1|\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} = 0, D \neq 1) = 0$, and $P(Y^0_t = 1|\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} = 0, D = 0) = 1$,$P(Y^0_t = 1|\overline{Y}^1_{t-1} = 0, \overline{Y}^0_{t-1} = 0, D \neq 0) = 1$.

$$\begin{aligned}(A15) <& \frac{P(D=1)P(\overline{Y}_{t-1}=0|D=1)P(Y_t=1|\overline{Y}_{t-1}=0, D=1)}{P(D=1)P(\overline{Y}_{t-1}=0|D=1) + P(D=0)p^0_{t-1}(0,0) + P(D=0)p^0_{t-1}(0,\neq 0)} \\ &+ \frac{[P(D=0)p^0_{t-1}(0,0) + P(D=0)p^0_{t-1}(0,\neq 0)][1 - P(Y_t=1|\overline{Y}_{t-1}\neq 0, D=0)]}{P(D=1)P(\overline{Y}_{t-1}=0|D=1) + P(D=0)p^0_{t-1}(0,0) + P(D=0)p^0_{t-1}(0,\neq 0)}\end{aligned}$$

Monotonicity in $p^0_{t-1}(0,0)$ and $p^0_{t-1}(0,\neq 0)$ depends on the sign of $1 - P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0) - P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 1)$. If $1 - P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0) - P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 1) > 0$, the upper bound is increasing in both $p^0_{t-1}(0,0)$ and $p^0_{t-1}(0,\neq 0)$. Let them equal to 1, get:

$$\begin{aligned}UB_t =& \frac{P(D=1)P(\overline{Y}_{t-1}=0|D=1)P(Y_t=1|\overline{Y}_{t-1}=0, D=1)}{P(D=1)P(\overline{Y}_{t-1}=0|D=1) + P(D=0)} \\ &+ \frac{P(D=0)(1 - P(Y_t=1|\overline{Y}_{t-1}\neq 0, D=0))}{P(D=1)P(\overline{Y}_{t-1}=0|D=1) + P(D=0)}\end{aligned} \quad \text{(C.1.5)}$$

When $1 - P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0) - P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 1) < 0$, C.1.5 is decreasing in both $p^0_{t-1}(0,0)$ and $p^0_{t-1}(0,\neq 0)$. Let them equal to 0, get

$$UB_t = P(\overline{Y}_t = 1|\overline{Y}_{t-1} = 0, D = 0) \quad \text{(C.1.6)}$$

Combine these two upper bounds together, we get:

$$UB_t = \frac{P(D=1)P(\overline{Y}_{t-1}=0|D=1)P(Y_t=1|\overline{Y}_{t-1}=0,D=1)}{P(D=1)P(\overline{Y}_{t-1}=0|D=1)+P(D=0)} + \frac{P(D=0)max\{1-min\{P(Y_t=1|\overline{Y}_{t-1}=0,D=0),P(Y_t=1|\overline{Y}_{t-1}\neq 0,D=0)\}\}}{P(D=1)P(\overline{Y}_{t-1}=0|D=1)+P(D=0)} \tag{C.1.7}$$

C.2 Proof of Theorem 3

Under conditional mean independence assumption and positive MTS assumption, we have

$$P(Y_t^0=1|\overline{Y}_{t-1}^1=0,\overline{Y}_{t-1}^0=0,D=1) > P(Y_t^0=1|\overline{Y}_{t-1}^1=0,\overline{Y}_{t-1}^0=0,D=0)$$

$$P(Y_t^0=1|\overline{Y}_{t-1}^1=0,\overline{Y}_{t-1}^0\neq 0,D=1) > P(Y_t^0=1|\overline{Y}_{t-1}^1=0,\overline{Y}_{t-1}^0\neq 0,D=0)$$

Plug the inequalities into C.1.3, if $P(Y_t=1|\overline{Y}_{t-1}=0,D=0) > P(Y_t=1|\overline{Y}_{t-1}\neq 0,D=0)$, we have:

$$(A15) <$$
$$\frac{P(D=0)p_{t-1}^0(0,\neq 0)[P(Y_t=1|\overline{Y}_{t-1}=0,D=1)-P(Y_t=1|\overline{Y}_{t-1}\neq 0,D=0)]}{P(D=1)P(\overline{Y}_{t-1}=0|D=1)+P(D=0)p_{t-1}^0(0,0)+P(D=0)p_{t-1}^0(0,\neq 0)}$$
$$+\frac{P(D=1)P(\overline{Y}_{t-1}=0|D=1)[P(Y_t=1|\overline{Y}_{t-1}=0,D=1)-P(Y_t=1|\overline{Y}_{t-1}\neq 0,D=0)]}{P(D=1)P(\overline{Y}_{t-1}=0|D=1)+P(D=0)p_{t-1}^0(0,0)+P(D=0)p_{t-1}^0(0,\neq 0)}$$
$$+\frac{p_{t-1}^0(0,0)P(D=0)[P(Y_t|\overline{Y}_{t-1}=0,D=1)-P(Y_t=1|\overline{Y}_{t-1}=0,D=0)]}{P(D=1)P(\overline{Y}_{t-1}=0|D=1)+P(D=0)p_{t-1}^0(0,0)+P(D=0)p_{t-1}^0(0,\neq 0)}$$

which is increasing in $p_{t-1}^0(0,\neq 0)$ and decreasing in $p_{t-1}^0(0,0)$, take $p_{t-1}^0(0,\neq 0) = P(\overline{Y}_{t-1}\neq 0|D=0)$ and $p_{t-1}^0(0,0)=0$, we have: $UB_t = P(Y_t=1|\overline{Y}_{t-1}=$

$0 = 1) - P(Y_t = 1|\overline{Y}_{t-1} \neq 0 = 0)$ Using same way when $P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0) < P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)$, we can get $UB_t = P(Y_t = 1|\overline{Y}_{t-1} = 0 = 1) - P(Y_t = 1|\overline{Y}_{t-1} = 0 = 0)$. Comining them together:

$$UB_t = P(Y_t = 1|\overline{Y}_{t-1} = 0 = 1) - min\{P(Y_t = 1|\overline{Y}_{t-1} = 0 = 0), P(Y_t = 1|\overline{Y}_{t-1} \neq 0 = 0)\}$$

When negative MTS is assumed, we have

$$P(Y_t^1 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 = 0, D = 0) > P(Y_t^1 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 = 0, D = 1)$$

$$P(Y_t^1 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 \neq 0, D = 0) > P(Y_t^1 = 1|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 \neq 0, D = 1)$$

Plug these inequalities into (A15), if $P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0) < P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)$, we have:

$$(A15) >$$
$$\frac{P(D = 0)p_{t-1}^0(0, \neq 0)[P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 1) - P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)]}{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1) + P(D = 0)p_{t-1}^0(0, 0) + P(D = 0)p_{t-1}^0(0, \neq 0)}$$
$$+ \frac{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1)[P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 1) - P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)]}{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1) + P(D = 0)p_{t-1}^0(0, 0) + P(D = 0)p_{t-1}^0(0, \neq 0)}$$
$$+ \frac{p_{t-1}^0(0, 0)P(D = 0)[P(Y_t|\overline{Y}_{t-1} = 0, D = 1) - P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0)]}{P(D = 1)P(\overline{Y}_{t-1} = 0|D = 1) + P(D = 0)p_{t-1}^0(0, 0) + P(D = 0)p_{t-1}^0(0, \neq 0)}$$

The right hand side of this ieuqality is the same as the uppder bound under $P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0) > P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)$. and it gives same value that $LB_t = P(Y_t = 1|\overline{Y}_{t-1} = 0 = 1) - P(Y_t = 1|\overline{Y}_{t-1} \neq 0 = 0)$, when $P(Y_t = 1|\overline{Y}_{t-1} = 0, D = 0) < P(Y_t = 1|\overline{Y}_{t-1} \neq 0, D = 0)$, the result is symmetric, thus

$$UB_t = P(Y_t = 1|\overline{Y}_{t-1} = 0 = 1) - max\{P(Y_t = 1|\overline{Y}_{t-1} = 0 = 0), P(Y_t = 1|\overline{Y}_{t-1} \neq 0 = 0)\}$$

C.3 Proof of Theorem 6

To identify difference in percentiles, first identify bounds on conditional CDF. Different from indentifying $ATETS_t$, here $P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1 = 0)$ and $P(Y_t^0 \leq \omega|\overline{Y}_{t-1}^1 = 0)$ needs to be identified separately.

$$
\begin{aligned}
P(Y_t^1 \leq & \omega|\overline{Y}_{t-1}^1 = 0) \\
&= \frac{P(D=1)P(\overline{Y}_{t-1}=0|D=1)P(Y_t \leq \omega|\overline{Y}_{t-1}=0, D=1)}{P(D=1)P(\overline{Y}_{t-1}=0|D=1) + P(D=0)p_{t-1}^0(0,0) + P(D=0)p_{t-1}^0(0,\neq 0)} \\
&+ \frac{P(D=0)p_{t-1}^0(0,0)P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 = 0, D=0)}{P(D=1)P(\overline{Y}_{t-1}=0|D=1) + P(D=0)p_{t-1}^0(0,0) + P(D=0)p_{t-1}^0(0,\neq 0)} \\
&+ \frac{P(D=0)p_{t-1}^0(0,\neq 0)P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1 = 0, \overline{Y}_{t-1}^0 \neq 0, D=0)}{P(D=1)P(\overline{Y}_{t-1}=0|D=1) + P(D=0)p_{t-1}^0(0,0) + P(D=0)p_{t-1}^0(0,\neq 0)}
\end{aligned} \tag{C.3.1}
$$

If $P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=1)$ first order stochastic dominates $P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=0)$,

$$
\begin{aligned}
&P(Y_t \leq \omega|\overline{Y}_{t-1}=0, D=1) \leq P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1=0) \\
&\leq \frac{P(D=1)P(\overline{Y}_{t-1}|D=1)P(Y_t=1|\overline{Y}_{t-1}=0|D=1) + P(D=0)}{P(D=0)P(\overline{Y}_{t-1}=0|D=1) + P(D=0)}
\end{aligned} \tag{C.3.2}
$$

The constant on right hand side cannot be used to give bounds on percentiles. Thus, under assumption that $P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=1)$ first order stochastic dominates $P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=0)$, we have a lower bound on on $P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1 = 0)$, which means we can obtain an upper bound on percentiles.

Using the same procedure on $P(Y_t^0 \leq \omega|\overline{Y}_{t-1}^1 = 0)$ and for cases when

$P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=1)$ is first order stochastic dominated by $P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=0)$, we have:

If $P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=1)$ first order stochastic dominates $P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=0)$

$$P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1 = 0) \geq P(Y_t \leq \omega|\overline{Y}_{t-1} = 0, D=1)$$

$$P(Y_t^0 \leq \omega|\overline{Y}_{t-1}^1 = 0) \leq max\{P(Y_t \leq \omega|\overline{Y}_{t-1} = 0, D=0), P(Y_t \leq \omega|\overline{Y}_{t-1} \neq 0, D=0)\}$$

If $P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=0)$ first order stochastic dominates $P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1, \overline{Y}_{t-1}^0, D=1)$

$$P(Y_t^1 \leq \omega|\overline{Y}_{t-1}^1 = 0) \leq P(Y_t \leq \omega|\overline{Y}_{t-1} = 0, D=1)$$

$$P(Y_t^0 \leq \omega|\overline{Y}_{t-1}^1 = 0) \geq min\{P(Y_t \leq \omega|\overline{Y}_{t-1} = 0, D=0), P(Y_t \leq \omega|\overline{Y}_{t-1} \neq 0, D=0)\}$$

Then we can use the way proposed by Manski and Sims (1994) to convert bounds on CDF to bounds on percentiles.

Bibliography

Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of econometrics*, 113(2): 231–263.

Abbring, Jaap H, and Gerard J Van den Berg. 2003. "The nonparametric identification of treatment effects in duration models." *Econometrica*, 71(5): 1491–1517.

Abrevaya, Jason. 2019. "Missing dependent variables in fixed-effects models." *Journal of econometrics*, 211(1): 151–165.

Abrevaya, Jason, and Stephen G Donald. 2017. "A GMM approach for dealing with missing data on regressors." *Review of Economics and Statistics*, 99(4): 657–662.

Andrea, Rotnitzky, Daniel Scharfstein, Ting-Li Su, and James Robins. 2001. "Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring." *Biometrics*, 57(1): 103–113.

Andrews, Donald W. K., and Gustavo Soares. 2010. "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection." *Econometrica*, 78(1): 119–157.

Andrews, Donald WK, and Marcia MA Schafgans. 1998. "Semiparametric estimation of the intercept of a sample selection model." *The Review of Economic Studies*, 65(3): 497–517.

Angrist, Joshua, and Ivan Fernandez-Val. 2010. "Extrapolate-ing: External validity and overidentification in the late framework." National Bureau of Economic Research.

Angrist, Joshua D, and Guido W Imbens. 1995*a*. "Identification and estimation of local average treatment effects." National Bureau of Economic Research.

Angrist, Joshua D, and Guido W Imbens. 1995*b*. "Two-stage least squares estimation of average causal effects in models with variable treatment intensity." *Journal of the American statistical Association*, 90(430): 431–442.

Bajari, Patrick, Han Hong, and Stephen P Ryan. 2010*a*. "Identification and estimation of a discrete game of complete information." *Econometrica*, 78(5): 1529–1568.

Bajari, Patrick, Han Hong, and Stephen P Ryan. 2010*b*. "Identification and estimation of a discrete game of complete information." *Econometrica*, 78(5): 1529–1568.

Balat, Jorge F, and Sukjin Han. 2018. "Multiple treatments with strategic interaction." *Available at SSRN 3182766*.

Balke, Alexander, and Judea Pearl. 1997. "Bounds on treatment effects from studies with imperfect compliance." *Journal of the American Statistical Association*, 92(439): 1171–1176.

Barnwell, Jean-Louis, and Saraswata Chaudhuri. 2018. "Efficient estimation in sub and full populations with monotonically missing at random data." Technical report, McGill University.

Baum, Christopher F, Mark E Schaffer, and Steven Stillman. 2003. "Instrumental variables and GMM: Estimation and testing." *The Stata Journal*, 3(1): 1–31.

Beresteanu, Arie, Ilya Molchanov, and Francesca Molinari. 2011. "Sharp identification regions in models with convex moment predictions." *Econometrica*, 79(6): 1785–1821.

Berry, Steven T. 1992. "Estimation of a Model of Entry in the Airline Industry." *Econometrica: Journal of the Econometric Society*, 889–917.

Bertanha, Marinho, and Guido W Imbens. 2019. "External validity in fuzzy regression discontinuity designs." *Journal of Business & Economic Statistics*, 1–39.

Bhattacharya, Jay, Azeem M Shaikh, and Edward Vytlacil. 2008*a*. "Treatment effect bounds under monotonicity assumptions: an application to Swan-Ganz catheterization." *American Economic Review*, 98(2): 351–56.

Bhattacharya, Jay, Azeem M Shaikh, and Edward Vytlacil. 2008*b*. "Treatment effect bounds under monotonicity assumptions: An application to swan-ganz catheterization." *The American Economic Review*, 98(2): 351–356.

Bhattacharya, Jay, Azeem M Shaikh, and Edward Vytlacil. 2012. "Treatment effect bounds: An application to Swan–Ganz catheterization." *Journal of Econometrics*, 168(2): 223–243.

Björn, Paul A, Quang H Vuong, et al. 1984. "Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation."

Blundell, Richard, Amanda Gosling, Hidehiko Ichimura, and Costas Meghir. 2007. "Changes in the distribution of male and female wages accounting for employment composition using bounds." *Econometrica*, 75(2): 323–363.

Borenstein, Severin. 1989. "Hubs and High Fares: Dominance and Market Power in the US Airline Industry." *RAND Journal of Economics*, 20: 344–365.

Bresnahan, Timothy F, and Peter C Reiss. 1990. "Entry in monopoly market." *The Review of Economic Studies*, 57(4): 531–553.

Bresnahan, Timothy F, and Peter C Reiss. 1991. "Entry and competition in concentrated markets." *Journal of Political Economy*, 977–1009.

Breunig, Christoph, and Peter Haan. 2018. "Nonparametric Regression with Selectively Missing Covariates." *arXiv preprint arXiv:1810.00411*.

Brinch, Christian N, Magne Mogstad, and Matthew Wiswall. 2017*a*. "Beyond LATE with a discrete instrument." *Journal of Political Economy*, 125(4): 985–1039.

Brinch, C, Magne Mogstad, and Matthew Wiswall. 2017*b*. "Beyond LATE with a discrete instrument." *Journal of Political Economy*, 125(4): 985–1039.

Carpenter, James R, Michael G Kenward, and Stijn Vansteelandt. 2006. "A comparison of multiple imputation and doubly robust estimation for analyses with missing data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3): 571–584.

Cattaneo, Matias D. 2010. "Efficient semiparametric estimation of multivalued treatment effects under ignorability." *Journal of Econometrics*, 155(2): 138–154.

Chaudhuri, Saraswata. 2020. "On Efficiency Gains From Multiple Incomplete Subsamples." *Econometric Theory*, 36(3): 488–525.

Chaudhuri, Saraswata, and David K Guilkey. 2016. "GMM with multiple missing variables." *Journal of Applied Econometrics*, 31(4): 678–706.

Chay, Kenneth Y., and Michael Greenstone. 2003. "The Impact of Air Pollution on Infant mortality: Evidence from Geographic Variation in Pollu-

tion Shocks Induced by a Recession." *The Quarterly Journal of Economics*, 1121–1167.

Cheng, Philip E. 1994. "Nonparametric estimation of mean functionals with data missing at random." *Journal of the American statistical association*, 89(425): 81–87.

Chen, Xiaohong. 2007. "Large sample sieve estimation of semi-nonparametric models." *Handbook of econometrics*, 6: 5549–5632.

Chen, Xiaohong, and Timothy Christensen. 2015. "Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation."

Chen, Xiaohong, Elie T Tamer, and Alexander Torgovitsky. 2011. "Sensitivity analysis in semiparametric likelihood models."

Chen, Xiaohong, Han Hong, Alessandro Tarozzi, et al. 2008. "Semiparametric efficiency in GMM models with auxiliary data." *The Annals of Statistics*, 36(2): 808–843.

Chen, Xiaoyan, Jieqing Tan, Zhi Liu, and Jin Xie. 2017. "Approximation of functions by a new family of generalized Bernstein operators." *Journal of Mathematical Analysis and Applications*, 450(1): 244–261.

Chernozhukov, Victor, and Christian Hansen. 2005. "An IV model of quantile treatment effects." *Econometrica*, 73(1): 245–261.

Chernozhukov, Victor, Han Hong, and Elie Tamer. 2007. "Estimation and confidence regions for parameter sets in econometric models 1." *Econometrica*, 75(5): 1243–1284.

Chesher, A, and A Rosen. 2012*a*. "Simultaneous Equations Models for Discrete Outcomes, Coherence, Completeness." and Identification, mimeo.

Chesher, Andrew. 2005. "Nonparametric identification under discrete variation." *Econometrica*, 73(5): 1525–1550.

Chesher, Andrew. 2010. "Instrumental variable models for discrete outcomes." *Econometrica*, 78(2): 575–601.

Chesher, Andrew, and Adam Rosen. 2012*b*. "Simultaneous equations models for discrete outcomes: coherence, completeness, and identification." *CeMMAP working paper, Centre for Microdata Methods and Practice*.

Chesher, Andrew, and Adam Rosen. 2014. "Generalized instrumental variable models." cemmap working paper, Centre for Microdata Methods and Practice.

Chesher, Andrew, and Adam Rosen. 2017. "Generalized instrumental variable models." *Econometrica, forthcoming*.

Chiburis, Richard C. 2010. "Semiparametric bounds on treatment effects." *Journal of Econometrics*, 159(2): 267–275.

Ciliberto, Federico, and Elie Tamer. 2009*a*. "Market structure and multiple equilibria in airline markets." *Econometrica*, 77(6): 1791–1828.

Ciliberto, Federico, and Elie Tamer. 2009*b*. "Market structure and multiple equilibria in airline markets." *Econometrica*, 77(6): 1791–1828.

Ciliberto, Federico, Charles Murry, and Elie Tamer. 2018. "Market Structure and Competition in Airline Markets." *University of Virginia, Penn State University, Harvard University.*

Coolidge, Julian L. 1949. "The story of the binomial theorem." *The American Mathematical Monthly*, 56(3): 147–157.

Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg. 2016. "From LATE to MTE: Alternative methods for the evaluation of policy interventions." *Labour Economics*, 41: 47–60.

Cunha, Flavio, James Heckman, and Salvador Navarro. 2005. "Separating uncertainty from heterogeneity in life cycle earnings." *oxford Economic papers*, 57(2): 191–261.

Deb, Rahul, Yuichi Kitamura, John Kim-Ho Quah, and Jörg Stoye. 2017. "Revealed price preference: Theory and stochastic testing."

Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii. 2019. "From local to global: External validity in a fertility natural experiment." *Journal of Business & Economic Statistics*, 1–27.

de Paula, Aureo. 2013. "Econometric analysis of games with multiple equilibria." *Annu. Rev. Econ.*, 5(1): 107–131.

de Paula, Áureo, and Xun Tang. 2012. "INFERENCE OF SIGNS OF INTERACTION EFFECTS IN SIMULTANEOUS GAMES WITH INCOMPLETE INFORMATION." *Econometrica*, 143–172.

Dickstein, Michael J., and Eduardo Morales. 2018. "What do Exporters Know?" *The Quarterly Journal of Economics*, 133(4): 1753—1801.

Dunlop, Dorothy D, Larry M Manheim, Jing Song, and Rowland W Chang. 2002. "Gender and ethnic/racial disparities in health care utilization among older adults." *The Journals of Gerontology Series B: Psychological sciences and social sciences*, 57(4): S221–S233.

Elbers, Chris, and Geert Ridder. 1982. "True and spurious duration dependence: The identifiability of the proportional hazard model." *The Review of Economic Studies*, 49(3): 403–409.

Federal Aviation Administration. 2015. "Aviation Emissions, Impacts & Mitigation: A Primer." *FAA, Office of Environment and Energy.*

Feng, Qian. 2016. "Instrumental Variables Estimation with Missing Instruments." Mimeo.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P Newhouse, Heidi Allen, Katherine

Baicker, and Oregon Health Study Group. 2012. "The Oregon health insurance experiment: evidence from the first year." *The Quarterly journal of economics*, 127(3): 1057–1106.

Firpo, Sergio, and Geert Ridder. 2019. "Partial identification of the treatment effect distribution and its functionals." *Journal of Econometrics*, 213(1): 210–234.

Foster, Andrew, and Mark Rosenzweig. 2008. "Inequality and the sustainability of agricultural productivity growth: Groundwater and the Green Revolution in rural India."

Gad, Ahmed M. 2011. "A selection model for longitudinal data with non-ignorable non-monotone missing values." *Journal of Data Science*, 9(171-180).

Galichon, Alfred, and Marc Henry. 2011. "Set identification in models with multiple equilibria." *The Review of Economic Studies*, 78(4): 1264–1298.

Gentzkow, Matthew, Jesse M Shapiro, and Michael Sinkinson. 2011. "The effect of newspaper entry and exit on electoral politics." *The American Economic Review*, 101(7): 2980–3018.

Glynn, Adam N, and Kevin M Quinn. 2010. "An introduction to the augmented inverse propensity weighted estimator." *Political analysis*, 36–56.

Goolsbee, Austan, and Chad Syverson. 2008. "How do incumbents respond to the threat of entry? Evidence from the major airlines." *The Quarterly Journal of Economics*, 123(4): 1611–1633.

Groenwold, Rolf HH, A Rogier T Donders, Kit CB Roes, Frank E Harrell Jr, and Karel GM Moons. 2012. "Dealing with missing outcome data in randomized trials and observational studies." *American journal of epidemiology*, 175(3): 210–217.

Gunsilius, Florian. 2019. "Bounds in continuous instrumental variable models." *arXiv preprint arXiv:1910.09502*.

Han, Sukjin. 2019. "Identification in Nonparametric Models for Dynamic Treatment Effects." *UT Austin*.

Han, Sukjin. 2020*a*. "Nonparametric estimation of triangular simultaneous equations models under weak identification." *Quantitative Economics*, 11(1): 161–202.

Han, Sukjin. 2020*b*. "Optimal Dynamic Treatment Regimes and Partial Welfare Ordering." *arXiv preprint arXiv:1912.10014*.

Han, Sukjin, and Edward J Vytlacil. 2017. "Identification in a generalization of bivariate probit models with dummy endogenous regressors." *Journal of Econometrics*, 199(1): 63–73.

Han, Sukjin, and Sungwon Lee. 2019. "Estimation in a generalization of bivariate probit models with dummy endogenous regressors." *Journal of Applied Econometrics*, 34(6): 994–1015.

Heckman, James J, and Edward J Vytlacil. 1999. "Local instrumental variables and latent variable models for identifying and bounding treatment effects." *Proceedings of the national Academy of Sciences*, 96(8): 4730–4734.

Heckman, James J, and Edward J Vytlacil. 2007*a*. "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation." *Handbook of econometrics*, 6: 4779–4874.

Heckman, James J, and Edward J Vytlacil. 2007*b*. "Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments." *Handbook of econometrics*, 6: 4875–5143.

Heckman, James J, and Edward Vytlacil. 2005. "Structural equations, treatment effects, and econometric policy evaluation1." *Econometrica*, 73(3): 669–738.

Heckman, James J, and Salvador Navarro. 2007. "Dynamic discrete choice and dynamic treatment effects." *Journal of Econometrics*, 136(2): 341–396.

Heckman, James J, Sergio Urzua, and Edward Vytlacil. 2006. "Understanding instrumental variables in models with essential heterogeneity." *The Review of Economics and Statistics*, 88(3): 389–432.

Heckman, J, and R Pinto. 2018. "Unordered monotonicity." *Econometrica*, 86(1): 1–35.

Heitjan, Daniel F, and Srabashi Basu. 1996. "Distinguishing "missing at random" and "missing completely at random"." *The American Statistician*, 50(3): 207–213.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American Statistical Association*, 81(396): 945–960.

Horowitz, Joel L, and Charles F Manski. 2000. "Nonparametric analysis of randomized experiments with missing covariate and outcome data." *Journal of the American statistical Association*, 95(449): 77–84.

Hsieh, Yu-Wei, Xiaoxia Shi, and Matthew Shum. 2018. "Inference on estimators defined by mathematical programming." *Available at SSRN 3041040*.

Hurd, Michael D, and Kathleen McGarry. 1997. "Medical insurance and the use of health care services by the elderly." *Journal of Health Economics*, 16(2): 129–154.

Imbens, Guido W, and Joshua D Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

Joy, Kenneth I. 2000. "Bernstein polynomials." *On-Line Geometric Modeling Notes*, 13.

Jun, Sung Jae, Joris Pinkse, and Haiqing Xu. 2011. "Tighter bounds in triangular systems." *Journal of Econometrics*, 161(2): 122–128.

Kalai, Ehud. 2004. "Large robust games." *Econometrica*, 72(6): 1631–1665.

Kamat, Vishal. 2019. "Identification with latent choice sets: The case of the head start impact study." *arXiv preprint arXiv:1711.02048*.

Khan, Shakeeb, and Elie Tamer. 2010. "Irregular identification, support conditions, and inverse weight estimation." *Econometrica*, 78(6): 2021–2042.

Kitamura, Yuichi, and Jörg Stoye. 2018. "Nonparametric analysis of random utility models." *Econometrica*, 86(6): 1883–1909.

Kitamura, Yuichi, and Jörg Stoye. 2019. "Nonparametric Counterfactuals in Random Utility Models." *arXiv preprint arXiv:1902.08350*.

Kline, Brendan. 2015. "Identification of complete information games." *Journal of Econometrics*, 189(1): 117–131.

Kline, Brendan, and Elie Tamer. 2012. "Bounds for best response functions in binary games." *Journal of Econometrics*, 166(1): 92–105.

Kline, Patrick, and Christopher R Walters. 2019. "On Heckits, LATE, and numerical equivalence." *Econometrica*, 87(2): 677–696.

Knittel, Christopher R., Douglas Miller, and Nicholas J. Sanders. 2011. "Caution, drivers! Children present. Traffic, pollution, and infant health." *working paper*.

Kowalski, Amanda E. 2016. "Doing more when you're running LATE: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments." National Bureau of Economic Research.

Kowalski, Amanda E. 2020. "Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform." National Bureau of Economic Research.

Lee, Sokbae, and Bernard Salanié. 2018. "Identifying effects of multivalued treatments." *Econometrica, forthcoming*.

Li, Sophia, Joe Mazur, Yongjoon Park, James Roberts, Andrew Sweeting, and Jun Zhang. 2018. "Endogenous and Selective Service Choices After Airline Mergers." *working paper*.

Little, Roderick JA. 1992. "Regression with missing X's: a review." *Journal of the American statistical association*, 87(420): 1227–1237.

Little, Roderick JA, and Donald B Rubin. 2002. "Bayes and multiple imputation." *Statistical analysis with missing data*, 200–220.

Liu, Nianqing, Quang Vuong, Haiqing Xu, et al. 2013. "Rationalization and identification of discrete games with correlated types." Working paper.

Machado, Cecilia, Azeem Shaikh, and Edward Vytlacil. 2019. "Instrumental variables and the sign of the average treatment effect." *Journal of Econometrics*, 212: 522–555.

Magnac, Thierry, and David Thesmar. 2002. "Identifying dynamic discrete decision processes." *Econometrica*, 70(2): 801–816.

Majumdar, Rajeshwari. 2017. "On Affine and Conjugate Nonparametric Regression." *arXiv preprint arXiv:1710.06987*.

Manski, Charles F. 1990. "Nonparametric bounds on treatment effects." *The American Economic Review*, 80(2): 319–323.

Manski, Charles F. 1997. "Monotone treatment response." *Econometrica: Journal of the Econometric Society*, 1311–1334.

Manski, Charles F. 2007. "Partial identification of counterfactual choice probabilities." *International Economic Review*, 48(4): 1393–1410.

Manski, Charles F. 2013. "Identification of treatment response with social interactions." *The Econometrics Journal*, 16(1): S1–S23.

Manski, Charles F, and C Sims. 1994. "The selection problem." Vol. 1, 143–70.

Manski, Charles F, and John V Pepper. 2000*a*. "Monotone instrumental variables: With an application to the returns to schooling." *Econometrica*, 68(4): 997–1010.

Manski, Charles F, and John V Pepper. 2000*b*. "Monotone instrumental variables: With an application to the returns to schooling." *Econometrica*, 68(4): 997–1010.

Masten, Matthew A, and Alexandre Poirier. 2018. "Salvaging falsified instrumental variable models." *arXiv preprint arXiv:1812.11598*.

Menzel, Konrad. 2016. "Inference for games with many players." *The Review of Economic Studies*, 83: 306–337.

Meyer, Bruce D. 1996. "What have we learned from the Illinois reemployment bonus experiment?" *Journal of labor Economics*, 14(1): 26–51.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R Walters. 2019. "Identification of causal effects with multiple instruments: Problems and some solutions." National Bureau of Economic Research.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2017. "Using Instrumental Variables for Inference about Policy Relevant Treatment Effects." National Bureau of Economic Research.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using instrumental variables for inference about policy relevant treatment parameters." *Econometrica*, 86(5): 1589–1619.

Mourifié, Ismael. 2015. "Sharp bounds on treatment effects in a binary triangular system." *Journal of Econometrics*, 187(1): 74–81.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J Ganimian. 2019. "Disrupting education? Experimental evidence on technology-aided instruction in India." *American Economic Review*, 109(4): 1426–60.

Newey, KW, and Daniel McFadden. 1994. "Large sample estimation and hypothesis." *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, 2112–2245.

Newey, Whitney K. 1994. "The asymptotic variance of semiparametric estimators." *Econometrica: Journal of the Econometric Society*, 1349–1382.

Newey, Whitney K. 1997. "Convergence rates and asymptotic normality for series estimators." *Journal of econometrics*, 79(1): 147–168.

Newey, Whitney K. 2013. "Nonparametric instrumental variables estimation." *American Economic Review*, 103(3): 550–56.

Newey, Whitney K, and James L Powell. 2003. "Instrumental variable estimation of nonparametric models." *Econometrica*, 71(5): 1565–1578.

Pinto, Rodrigo. 2015. "Selection Bias in a Controlled Experiment: The Case of Moving to Opportunity." *University of Chicago*.

Possebom, Vitor. 2019. "Sharp Bounds for the Marginal Treatment Effect with Sample Selection." *arXiv preprint arXiv:1904.08522*.

Qi, Li, and Yanqing Sun. 2014. "Missing data approaches for probability regression models with missing outcomes with applications." *Journal of statistical distributions and applications*, 1(1): 23.

Robins, James M. 1997. "Non-response models for the analysis of non-monotone non-ignorable missing data." *Statistics in medicine*, 16(1): 21–37.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of regression coefficients when some regressors are not always observed." *Journal of the American statistical Association*, 89(427): 846–866.

Robins, James M, and Richard D Gill. 1997. "Non-response models for the analysis of non-monotone ignorable missing data." *Statistics in medicine*, 16(1): 39–56.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70(1): 41–55.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66(5): 688.

Scharfstein, Daniel O, Andrea Rotnitzky, and James M Robins. 1999. "Adjusting for nonignorable drop-out using semiparametric nonresponse models." *Journal of the American Statistical Association*, 94(448): 1096–1120.

Schlenker, Wolfram, and W Reed Walker. 2015. "Airports, air pollution, and contemporaneous health." *The Review of Economic Studies*, 83(2): 768–809.

Seaman, Shaun, John Galati, Dan Jackson, and John Carlin. 2013. "What Is Meant by" Missing at Random"?" *Statistical Science*, 257–268.

Seaman, Shaun R, and Ian R White. 2013. "Review of inverse probability weighting for dealing with missing data." *Statistical methods in medical research*, 22(3): 278–295.

Seaman, Shaun R, and Stijn Vansteelandt. 2018. "Introduction to double robust methods for incomplete data." *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33(2): 184.

Sekhri, Sheetal. 2014. "Wells, water, and welfare: the impact of access to groundwater on rural poverty and conflict." *American Economic Journal: Applied Economics*, 6(3): 76–102.

Shaikh, Azeem M, and Edward J Vytlacil. 2011. "Partial identification in triangular systems of equations with binary dependent variables." *Econometrica*, 79(3): 949–955.

Sun, BaoLuo, and Eric J Tchetgen Tchetgen. 2018. "On inverse probability weighting for nonmonotone missing at random data." *Journal of the American Statistical Association*, 113(521): 369–379.

Sun, BaoLuo, Neil J Perkins, Stephen R Cole, Ofer Harel, Emily M Mitchell, Enrique F Schisterman, and Eric J Tchetgen Tchetgen. 2018. "Inverse-probability-weighted estimation for monotone and nonmonotone missing data." *American journal of epidemiology*, 187(3): 585–591.

Tamer, Elie. 2003*a*. "Incomplete simultaneous discrete response model with multiple equilibria." *The Review of Economic Studies*, 70(1): 147–165.

Tamer, Elie. 2003*b*. "Incomplete simultaneous discrete response model with multiple equilibria." *The Review of Economic Studies*, 70(1): 147–165.

Tamer, Elie. 2010. "Partial identification in econometrics." *Annu. Rev. Econ.*, 2(1): 167–195.

Taubman, Sarah L, Heidi L Allen, Bill J Wright, Katherine Baicker, and Amy N Finkelstein. 2014. "Medicaid increases emergency-department use: evidence from Oregon's Health Insurance Experiment." *Science*, 343(6168): 263–268.

Tchetgen, Eric J Tchetgen, Linbo Wang, and BaoLuo Sun. 2018. "Discrete choice models for nonmonotone nonignorable missing data: Identification and inference." *Statistica Sinica*, 28(4): 2069–2088.

Tebaldi, Pietro, Alexander Torgovitsky, and Hanbin Yang. 2019. "Nonparametric estimates of demand in the california health insurance exchange." National Bureau of Economic Research.

Theil, Henri. 1971. "Principles of econometrics."

Torgovitsky, Alexander. 2019*a*. "Nonparametric Inference on State Dependence in Unemployment." *Econometrica*, 87: 1475–1505.

Torgovitsky, Alexander. 2019*b*. "Nonparametric inference on state dependence in unemployment." *Econometrica*, 87(5): 1475–1505.

Tsiatis, Anastasios. 2007. *Semiparametric theory and missing data.* Springer Science & Business Media.

Van Buuren, Stef. 2018. *Flexible imputation of missing data.* CRC press.

Vikström, Johan, Geert Ridder, and Martin Weidner. 2018. "Bounds on treatment effects on transitions." *Journal of Econometrics*.

Vuong, Quang, and Haiqing Xu. 2017. "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity." *Quantitative Economics*, 8(2): 589–610.

Vytlacil, Edward. 2002. "Independence, monotonicity, and latent index models: An equivalence result." *Econometrica*, 70(1): 331–341.

Vytlacil, Edward, and Nese Yildiz. 2007. "Dummy endogenous variables in weakly separable models." *Econometrica*, 75(3): 757–779.

Walker, Renee E, Christopher R Keane, and Jessica G Burke. 2010. "Disparities and access to healthy food in the United States: A review of food deserts literature." *Health & place*, 16(5): 876–884.

Wang, Lu, Andrea Rotnitzky, and Xihong Lin. 2010. "Nonparametric regression with missing outcomes using weighted kernel estimating equations." *Journal of the American Statistical Association*, 105(491): 1135–1146.

Wan, Yuanyuan, and Haiqing Xu. 2014. "Semiparametric identification of binary decision games of incomplete information with correlated private signals." *Journal of Econometrics*, 182(2): 235–246.

Woodbury, Stephen A, and Robert G Spiegelman. 1987. "Bonuses to workers and employers to reduce unemployment: Randomized trials in Illinois." *The American Economic Review*, 513–530.

Wooldridge, Jeffrey M. 2007. "Inverse probability weighted estimation for general missing data problems." *Journal of econometrics*, 141(2): 1281–1301.